LES PRECIS DU COFEB

N°22 • Décembre 2022

Par Fodiyé Bakary DOUCOURE, Ph.D.
Maître de Conférences Titulaire, Université Cheikh Anta DIOP, Dakar

STATISTIQUE DESCRIPTIVE

BCEAO
BANQUE CENTRALE DES ETATS DE L'AFRIQUE DE L'OUEST

COFEB
CENTRE OUEST AFRICAIN DE FORMATION ET D'ETUDES BANCAIRES

COFEB
Direction de la Recherche et des Partenariats

PRECIS DU COFEB

STATISTIQUE DESCRIPTIVE

Par Fodiyé Bakary DOUCOURE
Maître de Conférences Titulaire, Université Cheikh Anta DIOP, Dakar

TABLE DES MATIERES

TABLE DES MATIERES

TABLE DES MATIERES

PREAMBULE

Dans le cadre de sa mission d'éducation financière du public et de mise à jour permanente des connaissances, le Centre Ouest Africain de Formation et d'Etudes Bancaires (COFEB) a initié la série « Les Précis du COFEB ».

L'objectif principal visé consiste à vulgariser la connaissance économique à destination, non seulement des agents de la Banque Centrale des Etats de l'Afrique de l'Ouest (BCEAO) et des auditeurs du cycle diplômant du COFEB, mais aussi d'un public élargi intéressé par les thématiques abordées.

Il s'agit de documents synthétiques élaborés dans une démarche pédagogique et un langage accessible, permettant aux lecteurs non avertis de se faire une idée sur un thème précis et aux spécialistes de se rappeler des notions de base acquises lors de leur formation initiale. Les thèmes traités sont divers et couvrent aussi bien les fondamentaux économiques et financiers, les outils et techniques supports de l'analyse économique, que les questions dont l'intérêt est avéré.

Ce 22^{e} numéro de la série est intitulé « Statistique descriptive ».

Plus qu'un document informatif, c'est un ouvrage pédagogique qui tente de simplifier l'apprentissage d'une discipline comportant des formules et équations mathématiques pouvant s'avérer complexes, mais dont l'assimilation est rendue aisée eu égard à l'approche adoptée.

Ousmane SAMBA MAMADOU
Directeur Général du COFEB

MOT DU DIRECTEUR DE LA RECHERCHE ET DES PARTENARIATS

Dans l'accomplissement des missions d'une banque centrale, la plupart des décisions nécessite une analyse pointue de données chiffrées.

Nous présentons ainsi dans cet opus un aspect des sciences statistiques intitulé « Statistique descriptive ».

L'auteur, tenant compte des difficultés éventuelles à assimiler cette matière étroitement liée aux mathématiques, adopte une démarche didactique. Dans son approche d'enseignement, il procède par étape et présente les statistiques descriptives sous un angle « simplifié ». Il présente tout d'abord les définitions et notions de base, avant d'exposer les formules mathématiques, le tout étant alterné par des exercices d'application.

Ainsi, le COFEB, dans sa quête permanente de renforcement des capacités des agents de la BCEAO, n'a pas occulté la nécessité de vulgariser le savoir sur les statistiques. Il y va de l'aptitude de chaque agent à pouvoir prendre en charge les tâches dans les domaines métiers de la Banque, dont la réalisation requiert, dans la plupart des cas, la maîtrise de l'outil statistique.

Nous espérons que les utilisateurs de ce document y trouveront du plaisir et une pleine satisfaction, pour leurs divers besoins.

Bonne lecture !

Ndèye Amy NGOM SECK
Directeur de la Recherche et des Partenariats, COFEB

AVANT-PROPOS

Ecrit dans un langage relativement simple, par un universitaire, ce précis expose avec clarté et précision les méthodes de la statistique descriptive.

Le champ de ce précis se concentre sur la statistique dont l'objet est de fournir une description. Ni les lois de probabilités, ni la statistique inférentielle ne seront traitées.

Les logiciels statistiques permettent d'effectuer des calculs parfois complexes en des temps faibles ; aussi leur parfaite maîtrise nous paraît intimement liée aux objectifs de la statistique descriptive.

Il est à l'heure actuelle inconcevable de n'illustrer un cours de statistique descriptive qu'à l'aide des calculs effectués « à la main ».

Tous ces constats nous amènent à proposer un précis qui s'inscrit dans une nouvelle approche de l'enseignement de la statistique descriptive.

Ce livre possède l'avantage d'offrir simultanément une formation à la statistique descriptive et aux logiciels Eviews et Stata.

Cet ouvrage est le fruit de nos expériences. Nous l'avons conçu comme un tutorial, permettant au lecteur de trouver rapidement les réponses à ses interrogations et de franchir sans perdre de temps les difficultés auxquelles il pourrait se trouver confronté.

L'ouvrage comporte trois chapitres. Le premier présente les notions indispensables de la statistique descriptive et les représentations graphiques.

Le chapitre suivant donne les techniques utilisées pour l'analyse statistique des distributions à un caractère avec un accent mis sur les caractéristiques de tendance centrale, de dispersion, de forme et de concentration.

Nous verrons enfin dans un troisième chapitre les indices statistiques.

Ce document s'adresse ainsi principalement aux agents de la BCEAO.

Il sera également utile aux professionnels qui utilisent les techniques statistiques dans la mesure où chaque section est systématiquement illustrée par des exercices.

Enfin, je souhaite exprimer ma gratitude aux évaluateurs anonymes de la Direction de la Recherche qui ont accepté la très importante mais fastidieuse tâche de relire le manuscrit, et dont les remarques pertinentes m'ont permis plusieurs améliorations.

Selon l'expression consacrée, je demeure seul responsable des erreurs qui subsistent.

Fodiyé Bakary Doucouré

CHAPITRE 1 - VOCABULAIRE, DEFINITIONS ET REPRESENTATIONS GRAPHIQUES

Objectifs pédagogiques

A la fin de l'étude du chapitre 1, l'auditeur sera capable de :

1. mieux saisir l'importance de l'information numérique présentée sous diverses formes ;

2. préciser ce qu'on entend par population, unité statistique, échantillon, caractères, modalités, effectif, fréquence relative, fréquence relative cumulée ;

3. dépouiller les données et en établir les distributions des effectifs et des fréquences relatives ;

4. distinguer les différents types de caractères : quantitatif discret, quantitatif continu, qualitatif ordinal ; qualitatif nominal ;

5. tracer les principales représentations graphiques associées aux distributions, notamment le diagramme à secteurs, le diagramme à bandes, le diagramme en bâtons, le polygone des fréquences, la courbe en escaliers, l'histogramme et les courbes des fréquences relatives cumulées croissantes et décroissantes, la boîte à moustaches (box-plot).

1. Introduction

Les données ont été collectées, elles sont prêtes à être analysées.
Pour cela, la statistique a développé de nombreux outils de description de données.

La statistique descriptive a pour objet de résumer et de présenter l'information contenue dans les données collectées sur un groupe d'individus.

Les outils de la statistique descriptive varient selon que l'analyse concerne une, deux ou plusieurs variables : analyse univariée, bivariée et multivariée.

2. La statistique-une statistique-les statistiques

La statistique (le mot est employé au singulier avec l'article défini) évoque la science.

La statistique est la science qui a pour objet de recueillir un ensemble de données numériques relatives à tel ou tel phénomène et d'exploiter rationnellement ces données pour établir toutes relations de causalité par l'analyse et l'interprétation.

Il est bien connu de l'existence d'un grand nombre de définitions de la statistique.
Kendall ose même écrire que « parmi les thèmes à propos desquels les statisticiens ne sont pas d'accord se trouve la définition de leur science ».

Une statistique est une règle qui transforme un ensemble de données en une ou plusieurs valeurs numériques.

Une statistique est une fonction mesurable des observations.
Le terme est cette fois utilisé avec l'article indéfini.

Le chiffre d'affaires moyen d'un groupe d'entreprises est une statistique.

Les statistiques sont les données numériques qui interviennent pratiquement dans tous les domaines d'activité.
On peut citer, par exemple, les statistiques sur l'éducation et l'emploi.

3. Populations - Unités statistiques

Une **population** est l'ensemble des éléments auxquels se rapportent les données étudiées.

Tout élément de la population étudiée est appelé individu ou unité statistique, terme qui peut désigner aussi bien une personne (un manager) qu'un objet (une entreprise).

Ainsi la population Ω désigne l'ensemble de référence c'est à dire l'ensemble des unités statistiques observées.

On est souvent amené à ne considérer qu'un sous-ensemble E de la population Ω.

L'ensemble des individus de E est appelé échantillon et Card E s'appelle **taille de l'échantillon** et sera noté n.

4. Caractères et modalités

4.1 Caractères

Un caractère est un aspect particulier de l'individu auquel on s'intéresse.
On étudie les 4 caractères suivants relatifs aux 250 salariés d'une entreprise donnée :

- X : type de poste occupé ;
- Y : opinion sur l'entreprise ;
- Z : nombre d'années d'expérience ;
- T : rémunération mensuelle.

Un caractère peut prendre deux ou plusieurs modalités.

4.2 Modalités

Les modalités d'un caractère sont les différentes valeurs que peut prendre ce caractère sur l'ensemble de la population. Elles doivent former une partition, c'est à dire doivent être exhaustives et disjointes.
A chaque individu, on doit pouvoir associer une modalité et une seule.

Exemples de modalités

- le caractère « type de poste occupé » a 3 modalités : maintenance, production, administration.

- le caractère « opinion sur l'entreprise » a 5 modalités : très mauvaise, plutôt mauvaise, indifférente, plutôt bonne, très bonne.

- le caractère « nombre d'années d'expérience » peut avoir les modalités 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 ou plus.

- le caractère « rémunération mensuelle » peut avoir les modalités qui vont de 75 000 à 1 200 000 FCFA.

4.3 Les différents types de caractères

Un caractère peut être quantitatif ou qualitatif.

- s'il est qualitatif, il peut être nominal ou ordinal.
- s'il est quantitatif, il peut être discret ou continu.

La nature des modalités va donc conduire à distinguer quatre types de caractères : qualitatif nominal, qualitatif ordinal, quantitatif discret ou quantitatif continu.

4.3.1 Caractères qualitatifs

Un caractère est qualitatif s'il est lié à une observation ne pouvant pas faire l'objet d'une mesure.

Ses diverses modalités sont simplement constatées et repérées par un mot traduisant son état.
Les modalités d'un caractère qualitatif ne sont pas numériques.
On distingue deux types de caractères qualitatifs : d'une part le caractère qualitatif nominal, d'autre part le caractère qualitatif ordinal.

4.3.1.1 Caractère qualitatif nominal

Un caractère qualitatif est **nominal** si ses modalités ne peuvent pas être classées selon un ordre préétabli.
Le caractère qualitatif nominal est appelé caractère **qualitatif pur.**
Le type de poste occupé (modalités : maintenance, production, administration) est un caractère qualitatif nominal.

4.3.1.2 Caractère qualitatif ordinal

Un caractère qualitatif est **ordinal** si ses modalités peuvent se ranger selon un ordre précis.
L'opinion sur l'entreprise (modalités : très mauvaise, plutôt mauvaise, indifférente, plutôt bonne, très bonne) est un caractère qualitatif ordinal.
On peut considérer la modalité « très mauvaise » comme inférieure à la modalité « plutôt mauvaise » qui, elle-même, est inférieure à la modalité « plutôt bonne ».

4.3.2 Caractères quantitatifs

Un caractère est **quantitatif** si on peut le mesurer ou le compter : ses modalités sont numériques.
On peut effectuer des opérations algébriques (addition, soustraction, division, multiplication) sur un tel caractère.

On distingue deux types de caractères quantitatifs : d'une part le caractère quantitatif discret, d'autre part le caractère quantitatif continu.

4.3.2.1 Caractères quantitatifs discret

Un caractère quantitatif est **discret** (ou discontinu) si ses modalités prennent des valeurs entières. Ses modalités ne peuvent pas prendre des valeurs décimales.

Le nombre de personnes à charge est un caractère quantitatif discret ; il n'y a pas de valeurs possibles entre deux entiers consécutifs.
Par exemple le nombre de personnes à charge peut être 4 mais non 2,7.

4.3.2.2 Caractères quantitatifs continu

Un caractère quantitatif est dit **continu** s'il peut prendre toutes les valeurs possibles à l'intérieur d'un intervalle de IR. Ses modalités peuvent prendre des valeurs décimales.
La rémunération mensuelle est un caractère quantitatif continu.
Les variables regroupées en **classes** sont assimilées à des variables quantitatives continues pour le traitement statistique.

Comme les variables quantitatives continues possèdent un nombre de valeurs distinctes très important, on est amené pour plus de commodité à les regrouper en un certain nombre de classes. Une **classe** c_i est un intervalle de IR et s'écrit généralement sous la forme : $[b_i , b_{i+1}[$.

Les nombres b_i et b_{i+1} sont les bornes de la classe, b_i est la borne inférieure et b_{i+1} est la borne supérieure.

La différence $a_i = b_{i+1} - b_i$ s'appelle **amplitude** de la classe.

La valeur $\frac{b_i + b_{i+1}}{2}$ équidistante des deux bornes s'appelle **centre** de la classe.

Considérons la classe [100 000, 150 000[du caractère « rémunération mensuelle ».

- 100 000 est la borne inférieure ;
- 150 000 est la borne supérieure ;
- l'amplitude de la classe est 50 000 = 150 000 – 100 000 ;
- le centre de la classe est 125 000 = (100 000 + 150 000)/2

On appelle **effectif** d'une modalité X_i le nombre n_i d'individus observés ayant pris cette modalité

La **densité** de la classe c_i est la valeur $d_i = \frac{n_i}{a_i}$.

On utilise la densité quantité quand les classes n'ont pas la même amplitude.

Exercice 1 : Différents types de caractères

Les caractères suivants sont-ils qualitatifs ordinaux, qualitatifs nominaux, quantitatifs discrets ou quantitatifs continus ?

taille des villes, sexe, nombre d'enfants, état matrimonial, région habitée, transferts reçus, groupe sanguin, réaction à un vaccin, revenu, religion, ethnie, niveau de satisfaction, nombre de pièces du logement, type de logement, nombre d'écoles, diplôme, frais de scolarité, taux brut de scolarisation, nombre d'accidents de travail, valeur des exportations, catégorie socio-professionnelle, note obtenue à un examen, appréciation d'un produit électronique, nombre de personnes vaccinées, nombre de personnes vaccinées.

Pour les caractères qualitatifs ordinaux ou nominaux, il est demandé d'indiquer au moins deux modalités.

Solution

1. Les caractères qualitatifs ordinaux sont :

taille des villes (petite, moyenne, grande), réaction à un vaccin (nulle, légère, provoquer une ulcération, causer la mort), niveau de satisfaction (très insatisfait, insatisfait, satisfait, très satisfait), diplôme obtenu (aucun, baccalauréat, licence, master, doctorat), appréciation d'un produit électronique (très utile, assez utile, assez inutile, tout à fait inutile).

2. Les caractères qualitatifs nominaux sont :

sexe (masculin, féminin), état matrimonial (marié, divorcé, célibataire, veuf), région habitée (Dakar, Thiès, Saint-Louis,...), groupe sanguin (A+, A-, B+, B-, O+ , O-, AB+, AB-), religion (musulmane, chrétienne, animiste,…), ethnie (wolof, pular, sérère, diola, soninké,…), type de logement (maison, immeuble, case, baraque,...), catégorie socio-professionnelle (agriculteurs, cadres supérieurs, ouvriers, employés, chômeurs,...).

Les modalités des caractères sont entre parenthèses.

3. Les caractères quantitatifs discrets sont : nombre d'enfants, nombre de pièces du logement, nombre d'écoles, nombre d'accidents de travail, nombre de personnes vaccinées.

4. Les caractères quantitatifs continus sont : transferts reçus, revenu, frais de scolarité, taux brut de scolarisation, ventes de terres agricoles, valeur des exportations, note obtenue à un examen.

5. Représentations graphiques

« Un bon croquis vaut mieux qu'un long discours »
NAPOLEON

5.1 Graphiques associés à un caractère qualitatif

Pour représenter graphiquement les distributions statistiques relatives à un caractère qualitatif on utilise habituellement, soit des diagrammes à secteurs, soit des diagrammes à bandes.

Le diagramme à secteurs est particulièrement bien adapté aux variables qualitatives nominales, tandis que le diagramme à bandes est surtout utilisé pour les variables qualitatives ordinales.
Le diagramme à bandes permet de prendre en compte l'ordre qui existe entre les modalités de la variable.

Exercice 2 : Représentation graphique associée à un caractère qualitatif nominal

Le tableau suivant donne la catégorie professionnelle de 800 usagers d'un restaurant d'entreprise.

Catégorie professionnelle	Effectifs
Employé administratif	120
Cadre administratif	80
Ingénieur	136
Ouvrier	344
Agent de maîtrise	120
Total	800

1. Quelle est la population étudiée ? Quel est l'échantillon observé ? Quelle est l'unité statistique ?

2. Quel est le caractère observé ? Quelle est sa nature ? Quelles sont ses modalités ?

3. Faites une représentation graphique de ce tableau par un graphique à secteurs.

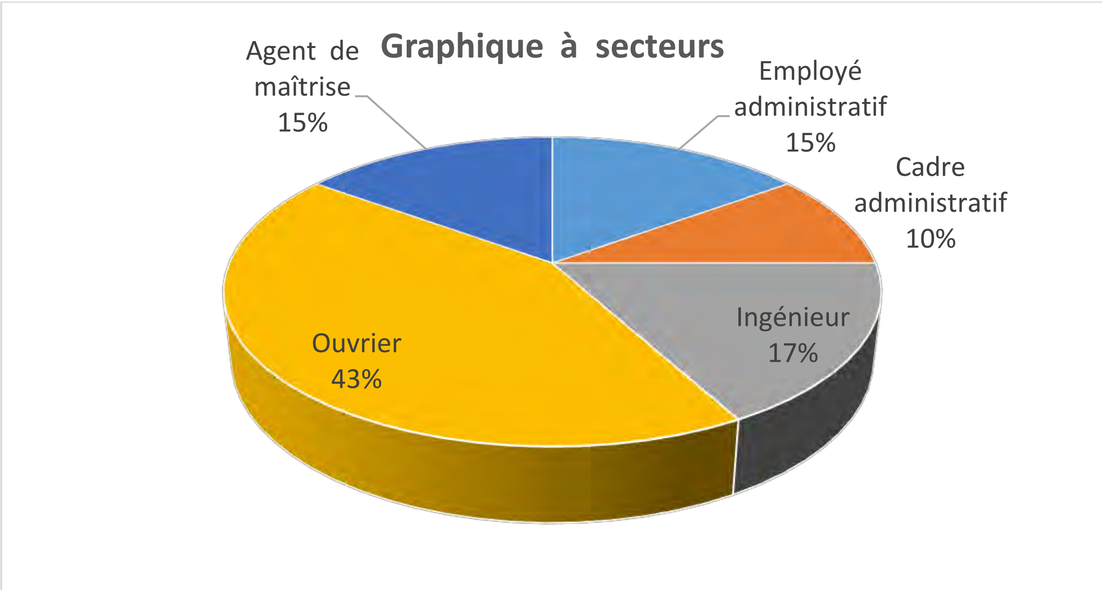
Solution

1. Population, échantillon et unité statistique
1.1 La population étudiée est l'ensemble des usagers.
1.2 L'échantillon observé est les 800 usagers.

1.3 L'unité statistique (ou l'individu) est un usager.

2. Caractère observé, nature et modalités
2.1 Le caractère observé est la catégorie professionnelle.
2.2 La catégorie professionnelle est un caractère qualitatif nominal.
2.3 Les modalités du caractère sont : employé administratif, cadre administratif, ingénieur, ouvrier et agent de maîtrise.

3. Graphique à secteurs



Exercice 3 : Représentation graphique associée à un caractère qualitatif ordinal

Le tableau ci-dessous donne la fréquence d'utilisation d'internet dans un échantillon de 189 entreprises.

Fréquence d'utilisation	Nombre d'entreprises
Très fréquemment	61
Fréquemment	71
Souvent	43
Rarement	13
Jamais	1
Total	189

1. Quelle est la population étudiée ? Quel est l'échantillon observé ? Quelle est l'unité statistique ?
2. Quel est le caractère observé ? Quelle est sa nature ?
3. Faites une représentation graphique de ce tableau par un diagramme à bandes.

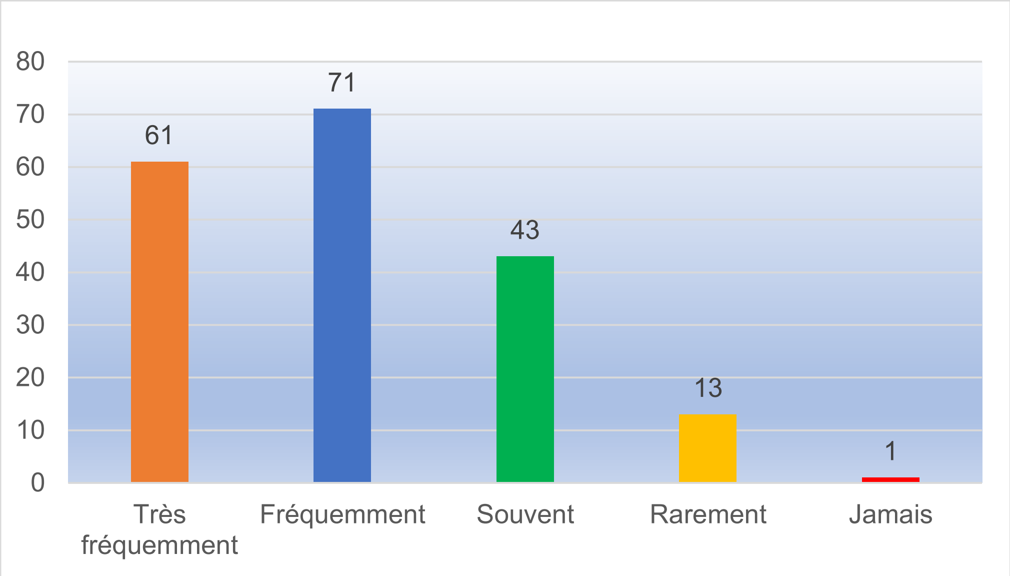
Solution

1. Population, échantillon et unité statistique
1.1 La population étudiée est l'ensemble des entreprises.
1.2 L'échantillon observé est les 189 entreprises.
1.3 L'unité statistique est une entreprise.

2. Caractère observé et nature
2.1 Le caractère observé est la fréquence d'utilisation d'internet.
2.2 La fréquence d'utilisation d'internet est un caractère qualitatif ordinal.

3. Diagramme à bandes verticales

Fréquence d'utilisation d'internet



5.2 Graphiques associés à un caractère quantitatif

5.2.1 Graphiques associés à un caractère quantitatif discret

5.2.1.1 Diagramme en bâtons

On porte sur l'axe des abscisses les valeurs discrètes du caractère, et sur l'axe des ordonnées les effectifs (ou fréquences) associés au caractère.

On trace des bâtons verticaux dont la longueur est proportionnelle aux effectifs (ou fréquences).

5.2.1.2 Polygone des fréquences

Le diagramme en bâtons étant construit, on peut définir le polygone des fréquences (ou des effectifs) qui a pour but de préciser l'évolution des effectifs.
On trace le polygone des fréquences en joignant les bouts des bâtons.

5.2.1.3 Courbe en escaliers

La courbe en escaliers (ou courbe cumulative) est la courbe représentative de la fonction F de la variable réelle x, telle que, pour toute valeur de x, la valeur prise par la fonction F, notée F(x), est égale à la proportion ou l'effectif des individus dont la valeur du caractère est strictement inférieure à x.

Exercice 4 : Représentations graphiques associées à un caractère quantitatif discret

Logiciels : Stata et Eviews

Une entreprise de services a relevé au cours des derniers mois, le nombre de plaintes par jour qui a été effectué à son service à la clientèle.

Les résultats obtenus sont indiqués dans le tableau ci-dessous :

0	0	0	1	1	2	0	0	0	0
3	1	0	2	2	0	0	1	0	2
1	1	0	3	1	0	0	0	1	1
0	0	2	4	1	0	0	1	1	1
1	1	0	0	1	3	2	0	1	1
1	0	3	0	1	1	0	1	1	1
0	0	1	0	1	1	2	2	0	2
1	2	1	3	4	2	1	0	2	2
0	0	1	1	1	2	1	2	0	3
1	0	2	2	2	0	0	0	0	0

1. Quelle est la population étudiée ? Quel est l'échantillon observé ? Quelle est l'unité statistique ?

2. Quel est le caractère observé ? Quelle est sa nature ? Quelles sont ses modalités ?

3. Dépouiller les renseignements qui précèdent et présenter les résultats du dépouillement sous la forme du tableau statistique ci-dessous :

Nombre de plaintes : x_i	0	1	2	3	4
Nombre de jours comportant x_i plaintes : n_i	…	…	…	6	…

4. Représenter graphiquement cette distribution par un diagramme en bâtons, un polygone des fréquences et une courbe en escaliers.

Solution

1. Population, échantillon et unité statistique
1.1 La population étudiée est l'ensemble des jours.
1.2 L'échantillon est les 100 jours.
1.3 L'unité statistique est un jour.

2. Caractère observé, nature et modalités
2.1 Le caractère observé est le nombre de plaintes par jour.
2.2 Le nombre de plaintes est un caractère quantitatif discret.
2.3 Les modalités du caractère sont : 0, 1, 2, 3 et 4.

3. Dépouillement des données avec Stata

```
tabulate  plaintes,  freq  plot
```

```
plaintes     |       freq.
------------+------------+-----------------------------------------------------
           0 |          39 |***************************************
           1 |          35 |***********************************
           2 |          18 |******************
           3 |           6 |******
           4 |           2 |**
------------+------------+-----------------------------------------------------
       Total |         100
```

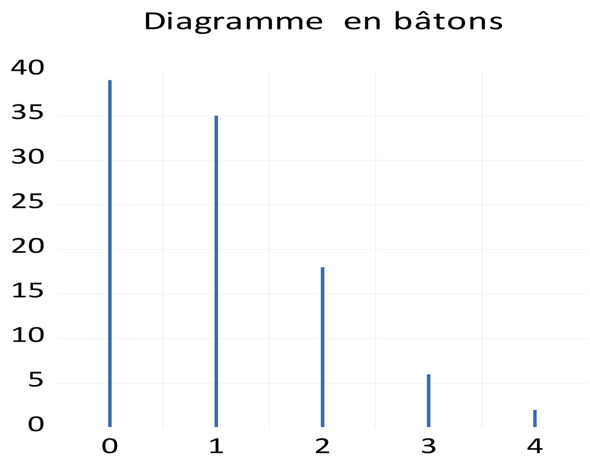
Le tableau statistique est le suivant :

Nombre de plaintes : x_i	0	1	2	3	4
Nombre de jours comportant x_i plaintes : n_i	39	35	18	6	2

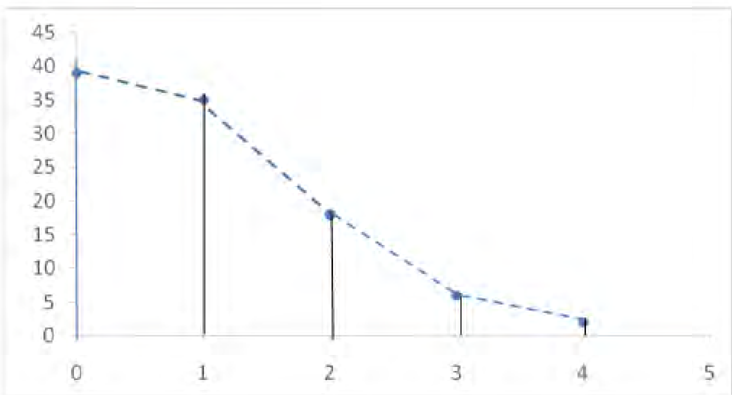
4. Représentations graphiques

4.1 Diagramme en bâtons avec Eviews

Quick → Graph → List of series : Saisir x → OK Specific : Choisir **Spike** → OK



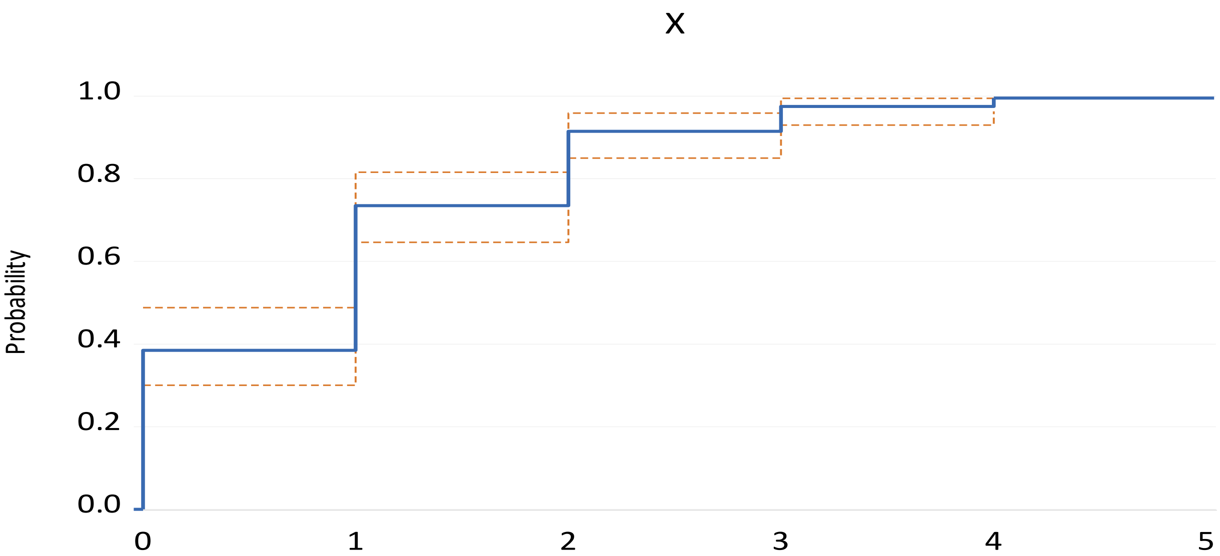
4.2 Polygone des fréquences



Le polygone des fréquences est la courbe en pointillés qui joint les bouts des bâtons.

4.3 Courbe en escaliers avec Eviews

Quick → Graph → List of series : Saisir x → OK Specific : Choisir Distribution → Distribution : Choisir **Empirical CDF** → OK



5.2.2 Graphiques associés à un caractère quantitatif continu

5.2.2.1 Histogramme

L'histogramme est la représentation graphique de la distribution des effectifs ou des fréquences d'une variable statistique.

Un histogramme est un ensemble de rectangles contigus, chaque rectangle, associé à chaque classe, a une surface proportionnelle à l'effectif de cette classe.

5.2.2.2 Polygone des fréquences (ou polygone des effectifs)

L'histogramme étant construit, on peut définir le polygone des fréquences (ou des effectifs) qui a pour but de préciser l'évolution des effectifs sur les différentes classes.

On trace le polygone des fréquences en joignant les milieux des segments supérieurs de chaque rectangle.

On peut éventuellement ajouter deux classes de même amplitude et d'effectif nul, de chaque côté de l'histogramme.
Ce polygone des fréquences a toujours une surface égale à la surface de l'histogramme.

5.2.2.3 Courbes des fréquences relatives cumulées croissantes et décroissantes

Comme pour les variables discrètes, la courbe des fréquences relatives cumulées croissantes est la représentation graphique de la fonction cumulative qui est égale à la proportion des observations pour lesquelles la variable statistique est inférieure à x.

Les observations étant groupées par classes ci, si ei représente l'extrémité supérieure de chaque classe, alors la courbe cumulative est la courbe qui passe par les points représentatifs de $F(e_i) = F_i$.

On peut également tracer une courbe associée aux fréquences relatives cumulées décroissantes.
La représentation graphique de sa fonction cumulative est $G = 1 – F$, elle représente le pourcentage des observations supérieures à x.
C'est une courbe monotone non croissante.

Exercice 5 : Représentations graphiques associées à un caractère quantitatif continu

Logiciel : Eviews

Dans une entreprise africaine, on a relevé les salaires mensuels de 40 employés.

Les valeurs (en milliers de FCFA) sont consignées dans le tableau suivant :

166	154	172	163	128	122	134	151	133	139
146	152	144	137	142	145	134	148	158	154
148	166	178	144	155	164	137	147	148	142
124	133	128	137	140	146	129	156	161	148

1. Quelle est la population étudiée ? Quel est l'échantillon observé ? Quelle est l'unité statistique ?

2. Quel est le caractère observé ? Quelle est sa nature ? Quelles sont ses modalités ?

3. Dépouiller les renseignements qui précèdent et présenter les résultats du dépouillement sous la forme du tableau statistique ci-dessous :

Classes	Effectifs
[120, 135[	
[135, 140[	
[140, 145[	
[145, 150[	8
[150, 155[	
[155, 160[	
[160, 170[	
[170, 180[	
Total	40

4. Tracer l'histogramme et le polygone des fréquences de cette distribution.

5. Tracer les courbes des fréquences relatives cumulées croissantes et décroissantes

6. Calculer les fréquences relatives cumulées croissantes et décroissantes.

7. Donner la signification des nombres inscrits à l'intersection de

7.1 la ligne classe [145,150[et la colonne «effectif n_i»

7.2 la ligne classe [145, 150[et la colonne «fréquence relative f_i»

7.3 la ligne classe [135, 140[et la colonne «fréquence relative cumulée croissante F_i ↗»

7.4 la ligne classe [155, 160[et la colonne «fréquence relative cumulée décroissante F_i ↘»

Solution

1.
1.1 La population étudiée est l'ensemble des employés de l'entreprise
1.2 L'échantillon observé est les 40 employés.
1.3 L'unité statistique est un employé.

2.
2.1 Le caractère observé est le salaire.
2.2 Le salaire est un caractère quantitatif continu.

3. Dépouillement des données avec Eviews

Cliquer sur Quick → Show → Objects to display in a single window : Saisir salaire → OK Cliquer sur View → One-Way Tabulation → OK

Le tableau statistique associé est :

salaire	n_i	f_i en %	F_i ↗ en %	F_i ↘ en %
[120, 135[	9	22,5	22,5	100
[135, 140[	4	10	32,5	77,5
[140, 145[	5	12,5	45	67,5
[145, 150[	8	20	65	55
[150, 155[	4	10	75	35
[155, 160[	3	7,5	82,5	25
[160, 170[	5	12,5	95	17,5
[170, 180[	2	5	100	5
Total	40	100		

Les fréquences relatives sont : $f_i = \frac{n_i}{n} \times 100$

Les fréquences relatives cumulées croissantes sont : F_i ↗

Les fréquences relatives cumulées décroissantes sont : F_i ↘

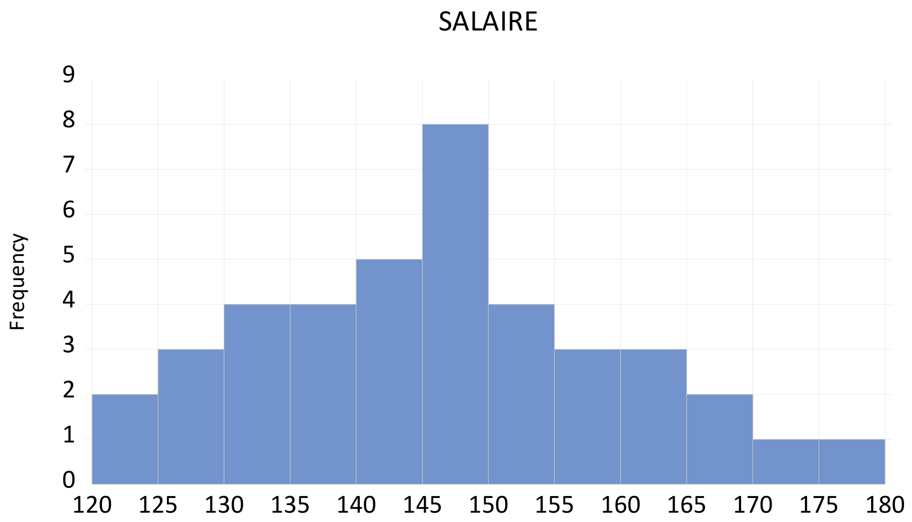
On a :

$$F_i \searrow \; = 100 - F_i \nearrow$$

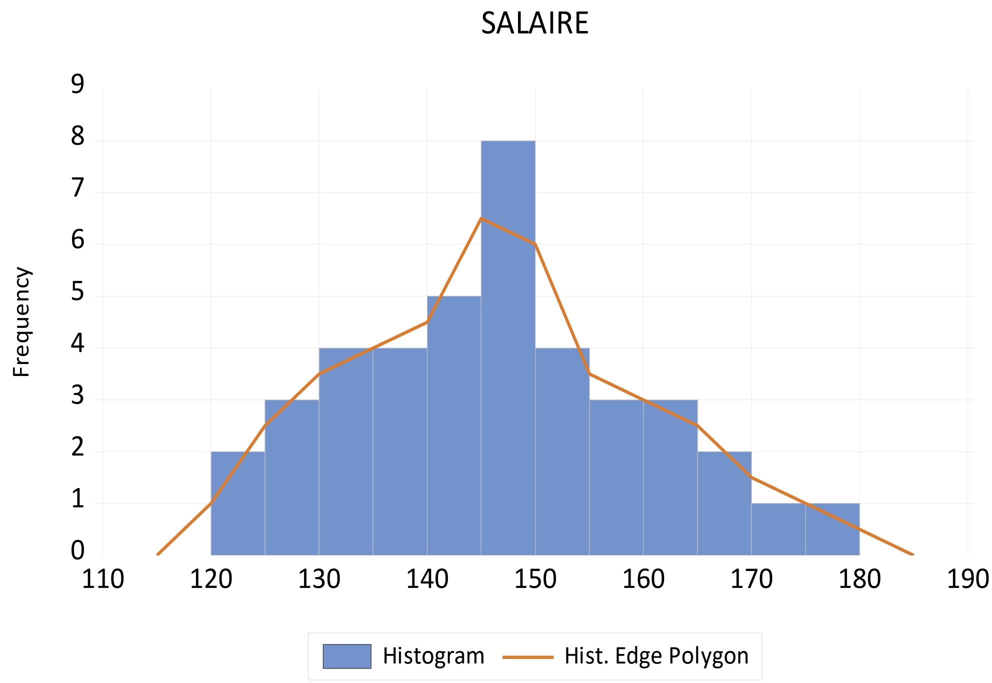
4. Histogramme et polygone des fréquences

4.1 Histogramme avec Eviews

Cliquer sur Quick → Graph → List of series : Saisir salaire → OK Specific : Choisir Distribution → OK



4.2 Polygone des fréquences avec Eviews

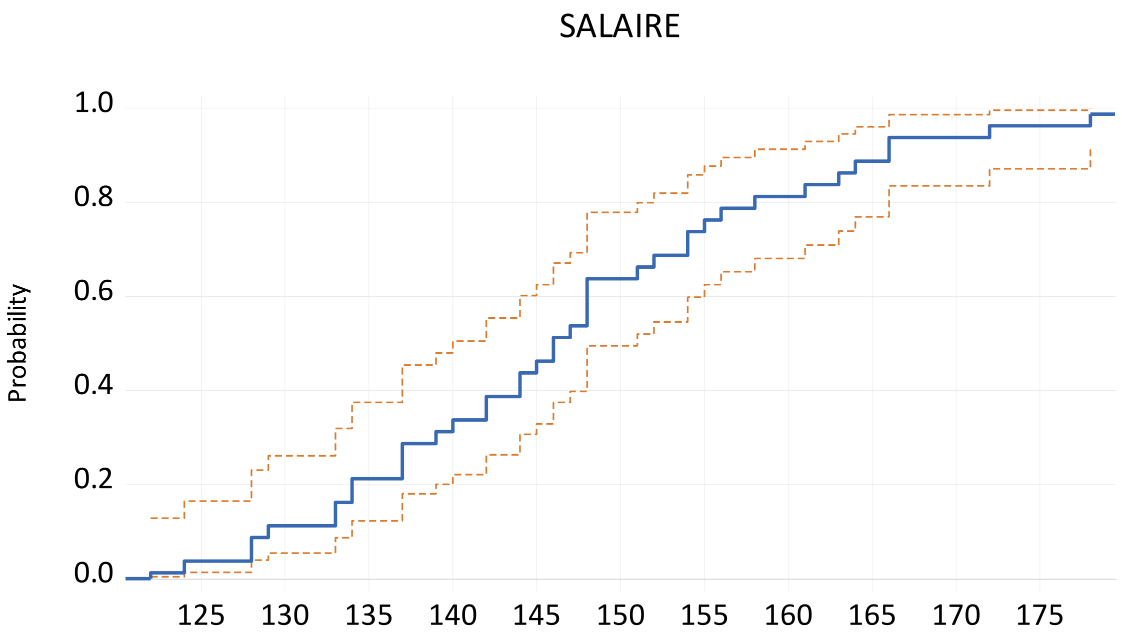


Le polygone des fréquences est la courbe en pointillés qui joint les milieux des segments supérieurs de chaque rectangle.

5. Courbe des fréquences relatives cumulées

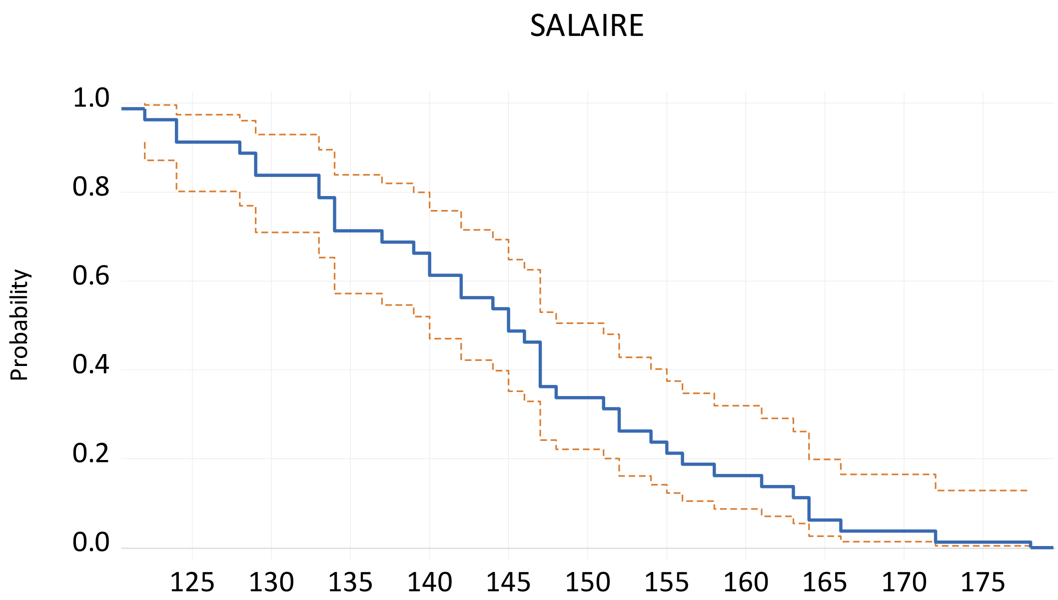
5.1 Courbe des fréquences relatives cumulées croissantes avec Eviews

Cliquer sur Quick → Graph → List of series : Saisir salaire → OK Specific : Choisir Distribution → Choisir **Empirical CDF** → OK



5.2 Courbe des fréquences relatives cumulées décroissantes avec Eviews

Cliquer sur Quick → Graph → List of series : Saisir salaire → OK Specific : Choisir Distribution → Choisir **Empirical Survivor** → OK



6. Les fréquences relatives cumulées croissantes et décroissantes sont calculées dans les tableaux de la question 3.

7.
7.1 Interprétation du nombre inscrit à l'intersection de la ligne classe [145,150[et de la colonne «effectif ni »
8 employés ont un salaire compris entre 145 000 et 150 000 FCFA.

7.2 Interprétation du nombre inscrit à l'intersection de la ligne classe [145, 150[et de la colonne «fréquence relative f_i»

20% des employés ont un salaire compris entre 145 000 et 150 000 FCFA.

7.3 Interprétation du nombre inscrit à l'intersection de la ligne classe [135, 140[et la colonne «fréquence relative cumulée croissante F_i ↗»

32,5% des employés ont un salaire strictement inférieur à 140 000 FCFA.

7.4 Interprétation du nombre inscrit à l'intersection de la ligne classe [155, 160[et la colonne «fréquence relative cumulée décroissante F_i ↘»

25% des employés ont un salaire supérieur ou égal à 155 000 FCFA.

CHAPITRE 2 – SERIE STATISTIQUE A UN CARACTERE

Objectifs pédagogiques

A la fin de l'étude du chapitre 2, l'auditeur sera capable de :

1. faire la distinction entre les caractéristiques de tendance centrale, de dispersion, de forme et de concentration ;

2. calculer avec les différentes formules qui sont présentées les moyennes, les quantiles, le mode et donner la signification concrète de chacune de ces mesures statistiques ;

3. calculer et interpréter la variance, l'écart-type, le coefficient de variation, l'étendue, les intervalles interquantiles et les écarts absolus moyens ;

4. préciser ce qu'on entend par asymétrie et aplatissement d'une distribution ;

5. interpréter les caractéristiques de concentration suivantes : courbe de concentration de Gini et indice de concentration de Gini ;

6. se servir des logiciels Stata et Eviews pour calculer les caractéristiques de tendance centrale, de dispersion, de forme et de concentration.

1. Caractéristiques de tendance centrale

« Si ce qui est simple est faux, ce qui est compliqué est inutile. »

PAUL VALERY.

Nous avons appris à ranger les données et à les présenter à l'aide de tableaux statistiques.

Nous avons également étudié les représentations graphiques qui constituent un moyen particulièrement adéquat de présentation des résultats.

Grâce à eux, on peut se faire une première idée de l'aspect d'une distribution statistique.

Cependant ces constatations visuelles demeurent imprécises et restent soumises aux dangers d'une appréciation synthétique forcément subjective.

Il faut donc trouver le moyen d'exprimer, autrement que par un commentaire de graphiques ou de tableaux, les éléments qui particularisent la série d'observations dont dispose le statisticien.

Pour cela, on utilise alors les caractéristiques de tendance centrale (ou de position) qui sont les moyennes, les quantiles et le mode.

1.1 Les moyennes

1.1.1 Moyenne arithmétique

La moyenne arithmétique d'une série statistique simple (x_i, i = 1 à n), est égale à la somme des valeurs observées, divisée par le nombre d'observations.

On la note généralement $\overline{x}$.

Ainsi

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

Dans le cas d'un tableau de distribution, on a

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$$

où $x_1, x_2, \ldots, x_p$ sont les valeurs observées (ou les centres des classes si la distribution est groupée), $n_1, n_2, \ldots, n_p$ sont les effectifs correspondants, $f_1, f_2, \ldots, f_p$ sont les fréquences correspondantes.

$$f_i = \frac{n_i}{n} \quad \text{et} \quad n = \sum_{i=1}^{p} n_i$$

Propriété de la moyenne arithmétique

Soit (x_i, i = 1 à n), une série statistique et (y_i, i = 1 à n), la série définie par $y_i = ax_i + b$ où a et b sont deux réels quelconques, alors :

$$\overline{y} = a\,\overline{x} + b$$

Démonstration

$$\begin{aligned}\overline{y} &= \frac{1}{n}\sum_{i=1}^{p} n_i y_i = \frac{1}{n}\sum_{i=1}^{p} n_i\left(a\,x_i + b\right)\\ &= a\,.\frac{1}{n}\sum_{i=1}^{p} n_i x_i + b\,.\frac{1}{n}\sum_{i=1}^{p} n_i\\ &= a\,.\,\overline{x} + b\end{aligned}$$

Cette formule de changement de variable permet de simplifier le calcul de la moyenne arithmétique dans certains cas.

1.1.2 Moyenne géométrique

La moyenne géométrique d'une série statistique positive (x_i, i = 1 à n), est la racine nième du produit des valeurs observées.

On la note généralement G.

Ainsi :

$$G = \sqrt[n]{x_1 x_2 x_3 \cdots x_n} = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}}$$

avec

$$\prod_{i=1}^{n} x_i = x_1 x_2 x_3 \cdots x_n$$

Dans le cas d'un tableau de distribution, on a :

$$G = \left[\prod_{i=1}^{p} {x_i}^{n_i}\right]^{\frac{1}{n}} = \prod_{i=1}^{p} {x_i}^{f_i}$$

1.1.3 Moyenne harmonique

La moyenne harmonique d'une série statistique strictement positive (x_i, i = 1 à n), est égale à l'inverse de la moyenne arithmétique des inverses des valeurs observées. On la note H.
Ainsi :

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

où

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

Dans le cas d'un tableau de distribution, on a :

$$H = \frac{\sum_{i=1}^{p} n_i}{\sum_{i=1}^{p} \frac{n_i}{x_i}} = \frac{1}{\sum_{i=1}^{p} \frac{f_i}{x_i}}$$

1.1.4 Moyenne quadratique

La moyenne quadratique d'une série statistique positive (x_i, i = 1 à n), est la racine carrée de la moyenne arithmétique des carrés des valeurs observées.

On la note Q.

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

Dans le cas d'un tableau de distribution, on a :

$$Q = \left[\frac{\sum_{i=1}^{p} n_i x_i^2}{\sum_{i=1}^{p} n_i} \right]^{\frac{1}{2}} = \left[\sum_{i=1}^{p} f_i x_i^2 \right]^{\frac{1}{2}}$$

Remarque : Soit une série pour laquelle les quatre moyennes définies ci-dessus existent, on a alors :

$$H < G < \overline{x} < Q$$

Exercice 6 : Calcul des moyennes d'une série statistique simple

Soit la série statistique simple :

1	2	5	7	10	13

Calculer les moyennes arithmétique, géométrique, harmonique, quadratique.

Solution

Partie 1

1. Moyenne arithmétique

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{6} [1 + 2 + 5 + 7 + 10 + 13] = \frac{1}{6} \times 38 = 6,333$$

2. Moyenne géométrique

$$G == [1 \times 2 \times 5 \times 7 \times 10 \times 13]^{1/6} = (9100)^{1/6} = 4,569$$

3. Moyenne harmonique

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{6}{1 + \frac{1}{2} + \frac{1}{5} + \frac{1}{7} + \frac{1}{10} + \frac{1}{13}} = 2{,}971$$

4. Moyenne quadratique

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{1}{6}(1^2 + 2^2 + 5^2 + 7^2 + 10^2 + 13^2)}$$

$$= \sqrt{58} = 7{,}616$$

On vérifie : H < G < $\overline{x}$ < Q.

Partie 2 : Calcul des moyennes avec Stata

```
ameans  x
```

```
    Variable |     Type             Obs        Mean
-------------+-----------------------------------------------
           x | Arithmetic             6       6.333333
             | Geometric              6       4.569201
             | Harmonic               6       2.97062
-------------------------------------------------------------
```

Le logiciel Stata indique les valeurs des moyennes arithmétique, géométrique et harmonique.

Calculons la moyenne géométrique.

1. Calcul du carré des valeurs observées

```
gen  xcarre  =  x*x
```

2. Calcul de la moyenne arithmétique du carré des valeurs observées

```
tabstat  xcarre, stat(mean)
```

```
variable     |      mean
-------------+----------
      xcarre |        58
------------------------
```

3. Calcul de la racine carrée de la moyenne arithmétique des carrés des valeurs observées

```
scalar  quadratic  =  sqrt(58)
```

4. Affichage à l'écran Stata de la moyenne quadratique

```
display quadratic
```

7.6157731

La moyenne géométrique vaut 7,616.

Exercice 7 : Moyennes géométrique et harmonique d'une série groupée

1. Le chiffre d'affaires d'un cabinet médical a connu une augmentation de 18% pendant les cinq premières années, de 24% par an pendant les quatre années suivantes, de 30% par an, pendant les trois années suivantes.
Calculer l'augmentation moyenne sur les douze années.

2. Un agent de change a acheté pour 1 500 000 FCFA d'actions à des cours différents, soit pour 800 000 FCFA au cours de 17 000 FCFA, 300 000 FCFA au cours de 15 000 FCFA et 400 000 FCFA au cours de 12 000 FCFA.
Calculer le cours moyen auquel lui reviennent ces actions.

Solution

1. L'augmentation moyenne a vérifié :

$$(1+a)^{12} = (1,18)^5 (1,24)^4 (1,30)^3$$

Soit

$$a = \sqrt[12]{(1,18)^5 (1,24)^4 (1,30)^3} - 1 = 0,229$$

L'augmentation moyenne sur les 12 ans est de 22,9%.

a est une moyenne géométrique pondérée par les augmentations.

2. Le cours moyen est :

$$\overline{C} = \frac{C_1 + C_2 + C_3}{n_1 + n_2 + n_1}$$

$$M_1 = 800\ 000\ ;\ M_2 = 300\ 000\ ;\ M_3 = 400\ 000$$

$$C_1 = 17\ 000\ ;\ C_2 = 15000\ ;\ C_3 = 12000\ ;\ M_i = n_i \times C_i$$

$$\overline{C} = \frac{M_1 + M_2 + M_3}{\frac{M_1}{C_1} + \frac{M_2}{C_2} + \frac{M_3}{C_3}} = \frac{800\ 000 + 300\ 000 + 400\ 000}{\frac{800\ 000}{17\ 000} + \frac{300\ 000}{15000} + \frac{400\ 000}{12000}}$$

$$= 14\ 941,41$$

Le cours moyen est égal à 14 941,41 FCFA.
Le cours moyen est une moyenne harmonique pondérée.

1.2 Les quantiles

On appelle quantile d'ordre α%, et on note Q_α, la valeur x_i du caractère telle que α% des valeurs observées soient inférieures strictement à x_i.

Si F désigne la fonction « fréquences relatives cumulées croissantes » alors :

$$F(Q_\alpha) = \frac{\alpha}{100}$$

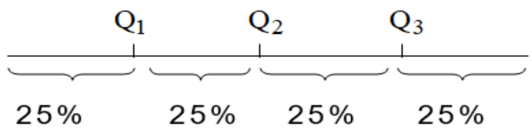
1.2.1 La médiane

Elle correspond au quantile d'ordre 50%. C'est donc la valeur du caractère étudié telle qu'il y ait autant d'observations qui lui soient supérieures que d'observations qui lui soient inférieures.

La médiane partage donc la série des valeurs observées en deux séries de même taille.

1.2.2 Les quartiles

On a 3 quartiles (Q_1, Q_2, Q_3) qui partagent la série en quatre séries de même taille.



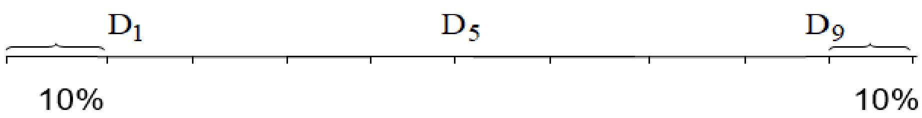
Q_1 est le premier quartile, c'est donc le quantile d'ordre 25%. Ce qui signifie que 25% des observations sont inférieures au premier quartile Q_1.

Q_2 est le deuxième quartile, c'est le quantile d'ordre 50%.
Q_2 est donc confondu avec la médiane. 50% des observations sont inférieures au deuxième quartile
Q_2 = Me

Q_3 est le troisième quartile, c'est le quantile d'ordre 75%. Ce qui signifie que 75% des observations sont inférieures au troisième quartile Q_3.

1.2.3 Les déciles

On a 9 déciles (D_1, D_2, D_3,…,D_9) qui partagent la série en 10 séries de même taille.



Le premier décile D_1 correspond au quantile d'ordre 10%.

Le cinquième décile correspond à la médiane.

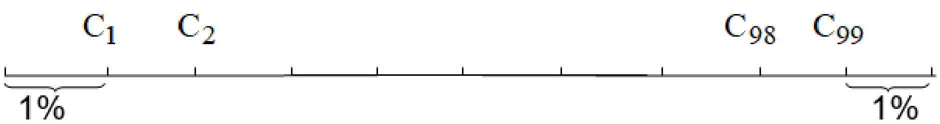
Le neuvième décile D_9 correspond au quantile d'ordre 90%.

10% des observations sont inférieures à D_1.

90% des observations sont inférieures au neuvième décile D_9.

1.2.4 Les centiles

On a 99 centiles (C_1, C_2, C_3,…,C_{99}) qui partagent la série en 100 séries de même taille.



Le premier centile C_1 correspond au quantile d'ordre 1%.

Le cinquantième centile correspond à la médiane.

Le quatre-vingt dix-neuvième centile C_{99} correspond au quantile d'ordre 99%.

1% des observations sont inférieures au premier centile C_1.

99% des observations sont inférieures à C_{99}.

Exercice 8 : Calcul des quartiles d'un caractère quantitatif discret

Logiciel : Stata

Une enquête effectuée auprès de 142 familles d'un pays africain a conduit à la distribution suivante, selon le nombre d'enfants.

0	1	2	3	4	5	6	7	8	9	10
0	1	2	3	4	5	6	7	8	9	10
0	1	2	3	4	5	6	7	8	9	10
0	1	2	3	4	5	6	7	4	9	10
0	1	2	3	4	5	9	7	4	4	10
0	1	2	9	10	5	9	7	6	4	10
2	0	2	9	10	5	4	10	6	4	6
2	0	2	9	10	5	4	10	6	4	6
3	3	2	6	10	8	4	10	6	4	6
3	3	2	3	4	8	4	10	6	5	4
3	9	2	3	4	3	6	1	3	5	4
3	9	4	3	4	5	6	1	6	5	4
3	4	3	6	1	1	6	3	3	4	

1. Dépouiller les renseignements qui précèdent et présenter les résultats du dépouillement sous la forme du tableau statistique suivant :

x_i	0	1	2	3	4	5
n_i	..	..	13	..	..	12

x_i	6	7	8	9	10
n_i	18	..	..	..	14

2. Calculer et interpréter les quartiles de cette distribution.

Solution

1. Dépouillement des données avec Stata

```
tabulate   x
```

```
  x          |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          8        5.63        5.63
          1 |         10        7.04       12.68
          2 |         13        9.15       21.83
          3 |         20       14.08       35.92
          4 |         25       17.61       53.52
          5 |         12        8.45       61.97
          6 |         18       12.68       74.65
          7 |          6        4.23       78.87
          8 |          5        3.52       82.39
          9 |         11        7.75       90.14
         10 |         14        9.86      100.00
------------+-----------------------------------
      Total |        142      100.00
```

Le tableau statistique obtenu est le suivant :

x_i	0	1	2	3	4	5
n_i	8	10	13	20	25	12

x_i	6	7	8	9	10
n_i	18	6	5	11	14

2. Calcul des quartiles avec Stata

```
tabstat  x , stat(q)
```

```
variable      |        p25        p50        p75
-------------+------------------------------
            x |          3          4          7
--------------------------------------------
```

Les quartiles Q_1, Q_2 et Q_3 valent respectivement 3, 4 et 7.

2.3 Interprétation des quartiles

2.3.1 Le premier quartile est égal à $Q_1 = 3$.

25% des familles ont moins de trois enfants.

2.3.2 Le deuxième quartile ou la médiane est égal à $Q_2 = 4$.

50% des familles ont moins de quatre enfants.

2.3.3 Le troisième quartile est égal à $Q_3 = 7$.

75% des familles ont moins de sept enfants.

1.3 Le mode

1.3.1 Définition

Le mode Mo d'une distribution statistique est sa valeur la plus fréquente. C'est la valeur du caractère qui correspond à l'effectif le plus grand ou à la fréquence la plus importante. Le mode permet ainsi de connaître la valeur la plus probable du caractère.

1.3.2 Détermination du mode

Deux cas se présentent pour sa détermination pratique.

1.3.2.1 Cas où la variable est quantitative discrète

Dans ce cas, le mode est défini avec précision. Il correspond à la valeur qui a l'effectif le plus élevé.

1.3.2.2 Cas où la variable est quantitative continue

Si la distribution est répartie en classes, le mode est indéterminé. Dans cette situation, on peut seulement définir la classe modale.

On peut distinguer deux cas :

Cas 1. Si les classes de la distribution sont d'amplitudes égales, la classe modale est celle d'effectif maximum.

Cas 2. Si les classes de la distribution n'ont pas la même amplitude, la classe modale est celle de densité maximum.

Remarque :

Une distribution peut avoir un ou plusieurs modes

a. Si une distribution statistique possède un seul mode, elle est dite unimodale ;

b. Si elle possède deux modes, elle est dite bimodale ;

c. Si elle possède plusieurs modes, elle est dite plurimodale.

Exercice 9 : Calcul du mode d'une variable quantitative discrète

Une enquête effectuée auprès de 142 familles d'un pays africain a conduit à la distribution suivante, selon le nombre d'enfants.

x_i	0	1	2	3	4	5	6	7	8
n_i	8	10	13	20	25	12	18	6	5

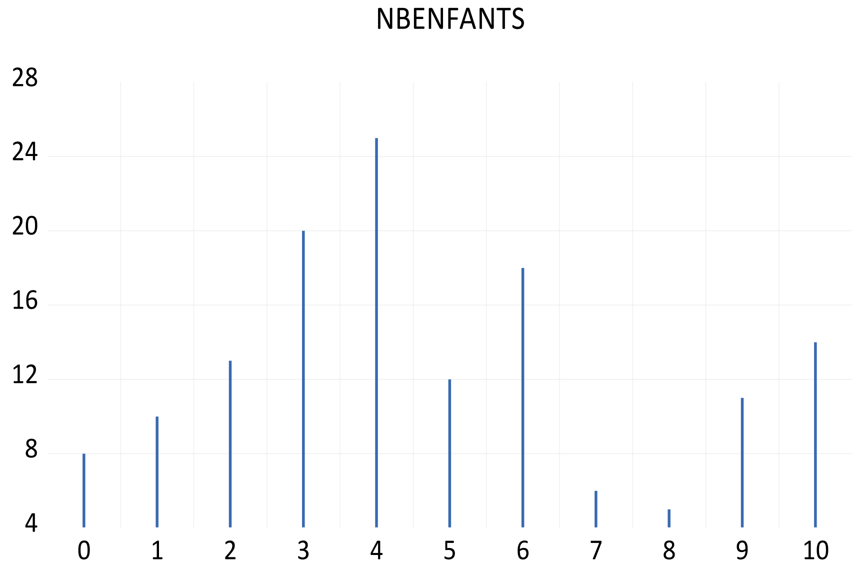
x_i	9	10
n_i	11	14

1. Tracer le diagramme en bâtons de cette distribution.

2. Calculer et interpréter le mode.

Solution

1. Diagramme en bâtons avec Eviews



2. Le mode est 4 car n_i est maximum pour $x_i = 4$.

Le nombre d'enfants le plus fréquent dans les familles est 4.

Calcul du mode avec le logiciel Stata

```
egen modex = mode(x) , minmode
```

```
list modex in 1
```

```
     +-------+
     | modex |
     |-------|
  1. |     4 |
     +-------+
```

Le mode est égal à 4.

Exercice 10 : Détermination de la classe modale d'une variable quantitative continue

Une enquête a été réalisée auprès de 210 employés d'un hôpital pour étudier la distribution des salaires nets mensuels.

Salaires en milliers de FCFA	Effectifs
[35 , 125 [	14
[125 , 150 [	38
[150, 220 [	60
[220 , 300 [	27
[300, 350 [	33
[350 , 425 [	20
[425 , 600 [	18

Déterminer la classe modale.

Solution

Le salaire est un caractère quantitatif continu, le mode est indéterminé.

Comme les classes n'ont pas la même amplitude, il nous faut calculer les densités.

Classes	n_i	a_i	d_i
[35 , 125 [	14	90	0,156
[125 , 150 [	38	25	**1,520**
[150, 220 [	60	70	0,857
[220 , 300 [	27	80	0,338
[300, 350 [	33	50	0,660
[350 , 425 [	20	75	0,267
[425 , 600 [	18	175	0,103
Total	**210**		

La classe modale est celle de densité maximale.

[125,150[est la classe modale avec une densité de 1,52.

Les salaires les plus fréquents sont compris entre 125 000 et 150 000 FCFA.

2. Caractéristiques de dispersion

« La diversité caractérise le mouvement. »

SALOUSTROS.

Les paramètres de dispersion sont des nombres qui mesurent la dispersion des valeurs observées autour d'un paramètre de position (moyenne, médiane,...).
Ces paramètres permettent de comparer des séries statistiques de même nature.

2.1 Les moments

2.1.1 Moments non centrés d'ordre r

On appelle moment non centré d'ordre r ($r \in$ IN) d'une variable X le nombre :

$$m_r(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^r = \sum_{i=1}^{p} f_i x_i{}^r$$

avec :

$$n = \sum_{i=1}^{p} n_i \quad ; \quad f_i = \frac{n_i}{n}$$

On a :

$$m_0(X) = \frac{1}{n}\sum_{i=1}^{p} n_i = 1 \ ; \ m_1(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \overline{x}$$

$$m_2(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2$$

centrés d'ordre r

n appelle moment centré d'ordre r (r… IN) le nombre :

$$\mu_r(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x})^r = \sum_{i=1}^{p} f_i (x_i - \overline{x})^r$$

avec :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i$$

On a :

$$\mu_0(X) = \frac{1}{n}\sum_{i=1}^{p} n_i = 1 \;\; ; \;\; \mu_1(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x}) = 0$$

$$\mu_2(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2$$

2.1.3 Changements d'origine et d'unité

Soit (x_i, i = 1 à n), une série statistique et (y_i, i = 1 à n), la série définie par $y_i = ax_i + b$, où a et b sont deux réels quelconques, alors :

$$\forall \; r \in \mathbb{N} \qquad \mu_r(Y) = a^r \, \mu_r(X)$$

Démonstration

$$y_i - \overline{y} = a\, x_i + b - (\, a\, \overline{x} + b)$$

$$= a\,(\, x_i - \overline{x}\,)$$

On a :

$$\mu_r(Y) = \frac{1}{n}\sum_{i=1}^{p} n_i (y_i - \overline{y})^r = \frac{1}{n}\sum_{i=1}^{p} n_i \left[a(x_i - \overline{x})\,\right]^r$$

$$= a^r \, \frac{1}{n}\sum_{i=1}^{p} n_i \, (x_i - \overline{x})^r = a^r \;\; \mu_r\,(X)$$

2.2 La variance et l'écart-type

2.2.1 La variance

On appelle variance d'une variable X son moment centré d'ordre 2 :

$$Var(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2 = \sum_{i=1}^{p} f_i (x_i - \overline{x})^2$$

Formule développée de la variance

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$$
$$= m_2 - m_1^2$$

Démonstration

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2$$
$$= \frac{1}{n}\sum_{i=1}^{p} n_i (x_i^2 - 2\,\bar{x}\,x_i + \bar{x}^2)$$

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - 2\,\bar{x}\,\frac{1}{n}\sum_{i=1}^{p} n_i x_i + \bar{x}^2\,\frac{1}{n}\sum_{i=1}^{p} n_i$$

$$= \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - 2\,\bar{x}.\bar{x} + \bar{x}^2.1 = \frac{1}{n}\sum_{i=1}^{p} n_i {x_i}^2 - \bar{x}^2$$

Propriété de la variance

Soit (x_i, i = 1 à n), une série statistique et (y_i, i = 1 à n), la série définie par $y_i = ax_i + b$, où a et b sont deux réels quelconques, alors :

$$\mathrm{Var}(Y) = a^2\,\mathrm{Var}(X)$$

2.2.2 L'écart-type

De tous les critères de dispersion, l'écart-type est certainement le plus utilisé.
L'écart-type d'une série est égal à la racine carrée de la variance :

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

L'écart type, noté σ (sigma) est une mesure de dispersion absolue, il s'exprime dans la même unité que les valeurs observées et mesure la dispersion autour de la moyenne $\bar{x}$.

Plus l'écart type est grand, plus la dispersion des observations autour de la moyenne est importante.

2.3 Le coefficient de variation

Pour faciliter les comparaisons entre séries, on utilise une mesure de dispersion relative appelée coefficient de variation.

Le coefficient de variation CV est le rapport de l'écart type σ à la moyenne $\bar{x}$:

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

C'est un nombre sans dimension et indépendant des unités choisies.

On l'utilise pour comparer par exemple des distributions de salaires dans différents pays.

Ainsi les salaires des enseignants ont pour coefficient de variation 0,45 alors que les salaires des médecins ont pour coefficient de variation 0,75 : le salaire est une variable relativement (c'est à dire compte tenu du rapport des moyennes) plus homogène chez les enseignants que chez les médecins.

Dans la pratique, on compare la valeur du coefficient de variation à 30%.

-- Une distribution est dite homogène si son coefficient de variation est inférieur à 30% ;

-- Une distribution sera considérée comme hétérogène si son coefficient de variation est supérieur ou égal à 30%.

Remarque :

Si le coefficient de variation dépasse 30%, alors les observations sont dispersées et il devient hasardeux d'interpréter la moyenne.

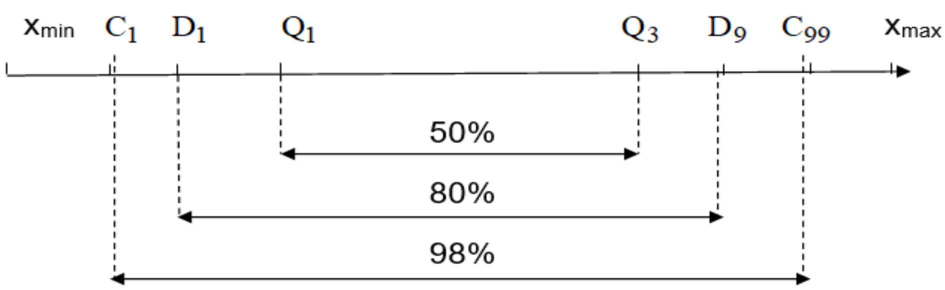
2.4 Les écarts interquantiles

2.4.1 Les intervalles interquantiles

L'intervalle interquartile est $[Q_1, Q_3[$, il contient 50% des observations.

L'intervalle interdécile est $[D_1, D_9[$, il contient 80% des observations.

L'intervalle intercentile est $[C_1, C_{99}[$, il contient 98% des observations.



2.4.2 Les écarts interquantiles

L'écart interquartile est le nombre $Q_3 - Q_1$.

L'écart interdécile est le nombre $D_9 - D_1$.

L'écart intercentile est le nombre $C_{99} - C_1$.

Ces écarts permettent de mesurer la dispersion de la série autour de la médiane.

2.5 L'étendue

C'est la valeur de dispersion la plus simple. L'étendue W est égale à la différence entre la valeur maximum observée et la valeur minimum observée.

On note :

$$W = X_{max} - X_{min}$$

L'étendue est un indice très élémentaire, il est très utilisé en contrôle de fabrication industrielle.

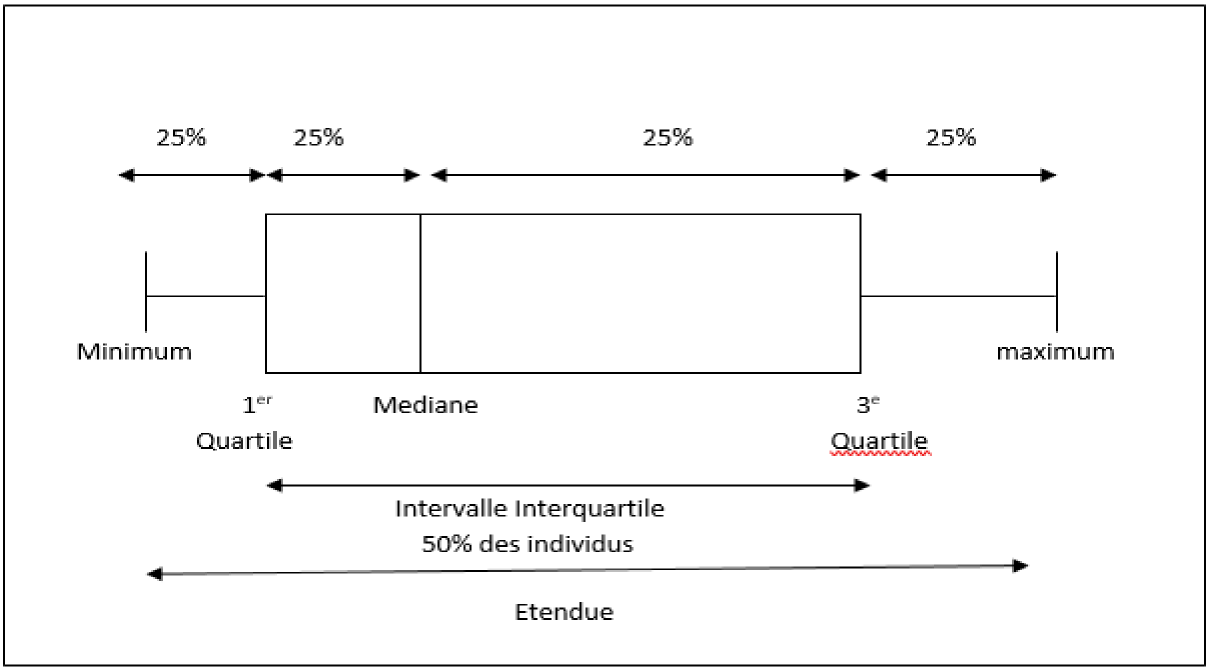
2.6 Diagramme en boîte et détection d'une variable aberrante

2.6.1 Diagramme en boîte

Il n'est pas toujours facile de comparer plusieurs distributions à l'aide d'histogramme ou de digramme en bâtons.

Il est alors intéressant d'avoir recours à des diagrammes en boîte appelés aussi boîtes de dispersion, aussi appelées boîtes à moustache ou encore boxplots.

Forme standard d'un diagramme en boîte



Une boîte de dispersion est une représentation graphique qui permet de visualiser l'étendue et l'intervalle interquartile.

La boîte centrale est délimitée par les valeurs des premier et troisième quartile ; elle a donc pour longueur l'écart interquartile.

La médiane est aussi représentée. Les segments de part et d'autre de la boîte relient l'extrémité de celle-ci aux valeurs minimum et maximum.

2.6.2 Détection de valeurs aberrantes

Les logiciels de statistique permettent aussi de visualiser les données aberrantes sur le diagramme en boîte.

Les données aberrantes sont alors considérées comme des valeurs extrêmes et représentées par des points isolés.

Une valeur aberrante est une donnée qui s'écarte de façon marquée de l'ensemble des données.

Tukey a donné une règle pratique pour identifier une valeur aberrante. Elle s'énonce comme suit :

Une donnée peut être appelée valeur aberrante si elle s'écarte d'une distance d'au moins 1,5 x ($Q_3 - Q_1$) au-dessus du troisième quartile ou en dessous du premier quartile.

Exercice 11 : Calcul des quartiles, détection d'une valeur aberrante, tracé du diagramme en boîte

Logiciel : Stata

On a relevé les données obtenues pour le test d'aptitude générale des 62 employés d'une entreprise.

93	104	93	79	78	112	107	100	105	102	107	107
119	94	87	113	98	86	124	93	99	97	83	95
99	98	77	101	104	138	97	74	99	85	93	98
84	110	102	75	104	100	84	101	82	85	85	92
86	101	70	108	89	68	123	63	86	62	90	77
94	96										

1. Déterminer les trois quartiles, les quatre plus petites valeurs et les quatre plus grandes de la série.

2. Déterminer l'écart interquartile.

3. Existe-t-il dans la série, des valeurs aberrantes ? Utiliser la règle de Tukey.

4. Tracer le diagramme en boîte (ou box-plot)

Solution

1. La série des données est notée x

```
sum  x , detail
```

```
                          x
-------------------------------------------------------------
      Percentiles      Smallest
 1%           62             62
 5%           70             63
10%           77             68       Obs                  62
25%           85             70       Sum of Wgt.          62

50%         95.5                      Mean           94.43548
                        Largest       Std. Dev.      14.54402
75%          102            119
90%          110            123       Variance       211.5286
95%          119            124       Skewness       .1797649
99%          138            138       Kurtosis       3.537089
```

1.1 Les quartiles Q_1, Q_2 et Q_3 valent respectivement 85 ; 95,5 et 102.

1.2 Les quatre plus petites valeurs (Smallest) sont 62, 63, 68 et 70.

La valeur minimale est 62.
1.3 Les quatre plus grandes valeurs (Largest) sont 119, 123, 124 et 138.

La valeur maximale est 138.

2. L'écart interquartile est égal à :

$$IQ = Q_3 - Q_1 = 102 - 85 = 17$$

3. Détection des valeurs aberrantes

3.1 Règle pratique pour détecter une valeur aberrante

Une valeur aberrante est une donnée qui s'écarte de façon marquée de l'ensemble des données.

Tukey a donné une règle pratique pour identifier une valeur aberrante.
Elle s'énonce comme suit :

Une donnée peut être appelée valeur aberrante si elle s'écarte d'une distance d'au moins $1{,}5 \times (Q_3 - Q_1)$ au-dessus du troisième quartile ou en dessous du premier quartile.

3.2 Application

On calcule d'abord

$$1{,}5 \times (Q_3 - Q_1) = 1{,}5 \times 17 = 25{,}5$$

Ainsi, on peut déclarer une donnée valeur aberrante

si elle est inférieure à $Q_1 - 1{,}5IQ = 85 - 25{,}5 = 59{,}5$

ou

si elle est supérieure $Q_3 + 1{,}5IQ = 102 + 25{,}5 = 127{,}5$

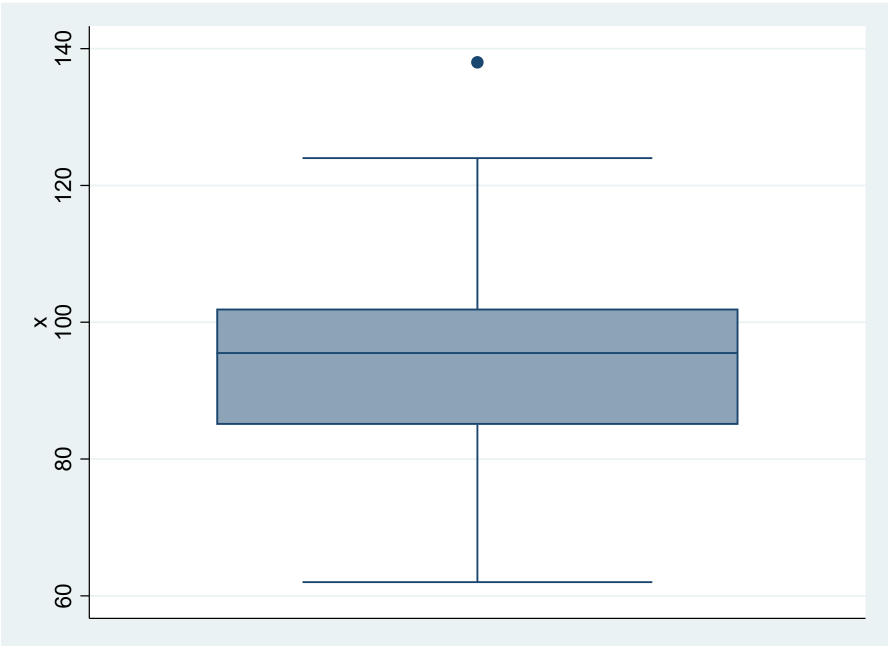
Les quatre plus petites valeurs sont 62, 63, 68 et 70.

Les quatre plus grandes valeurs sont 119, 123, 124 et 138.

Pour les données qui ont été obtenues, on constate qu'il y a une valeur extrême qui peut être déclarée valeur aberrante soit le résultat 138.

4. Diagramme en boîte (ou box-plot)

```
graph box x
```



3. Caractéristiques de forme

Les caractéristiques de forme permettent de préciser l'allure de la courbe des fréquences sans avoir recours à son tracé.

Les principales caractéristiques en ce qui concerne la forme sont : la dissymétrie (ou l'asymétrie) et l'aplatissement.

3.1 La dissymétrie

Il est toujours intéressant de savoir si une courbe de fréquence est symétrique ou non, et si elle ne l'est pas de mesurer sa dissymétrie. Il s'agit de mesurer son degré de dissymétrie.

L'asymétrie entraîne le fait suivant : les trois caractéristiques de tendance centrale, c'est-à-dire la moyenne arithmétique, le mode et la médiane ne sont plus confondus au centre de cette distribution, mais s'échelonnent dans un ordre différent selon que la distribution est étalée vers la gauche ou vers la droite.

On dispose de différents coefficients permettant de mesurer la dissymétrie.

3.1.1 Le coefficient d'asymétrie de Fisher

Il est défini par :

$$S = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

où

$$\mu_2 = \sigma^2 = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 \quad ; \quad \mu_3 = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^3$$

On doit distinguer trois cas.

Cas 1 : Si S > 0, la distribution est étalée vers la droite. On dit qu'elle a un biais positif. Dans cette situation on a :

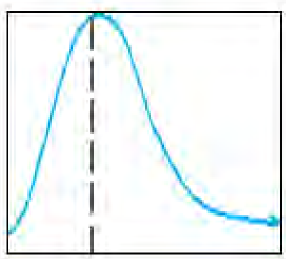
$$Mo < Me < \bar{x}$$

Cas 2 : Si S = 0, la distribution est symétrique. On dit qu'elle a un biais nul. Dans ce cas on a :

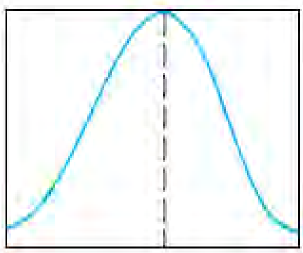
$$\bar{x} = Me = Mo$$

Cas 3 : Si S < 0, la distribution est étalée vers la gauche. On dit qu'elle a un biais négatif. Dans cette situation on a :

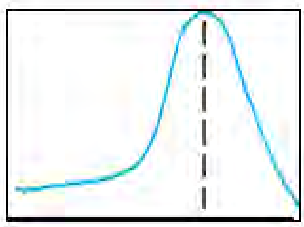
$$\bar{x} < Me < Mo$$



Distribution étalée vers la droite



Distribution symétrique



Distribution étalée vers la gauche

3.1.2 Le coefficient quartile de Yule

Il est défini par :

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

où Q_1 , Q_2 et Q_3 sont les trois quartiles.

On doit distinguer trois cas.

Cas 1 : Si $C_D > 0$, la distribution est étalée vers la droite.

Cas 2 : Si $C_D = 0$, la distribution est symétrique.

Cas 3 : Si $C_D < 0$, la distribution est étalée vers la gauche.

Exercice 12 : Calcul du coefficient de dissymétrie de Yule

Une enquête a été réalisée auprès de 210 employés d'un hôpital pour étudier la distribution des salaires nets mensuels.
Les quartiles de cette distribution sont :

$Q_1 = 150,491$; $Q_2 = 211,895$; $Q3 = 328,025$

Calculer et interpréter le coefficient de dissymétrie de Yule

Solution

Le coefficient de dissymétrie de Yule vaut :

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

$$= \frac{(328{,}025 - 211{,}895) - (211{,}895 - 150{,}491)}{328{,}025 - 150{,}491} = 0{,}308$$

Comme $C_D > 0$, la distribution des salaires est étalée vers la droite.
La distribution des salaires a un biais positif.

3.2 L'aplatissement

Une distribution statistique peut être plus ou moins aplatie selon qu'une proportion plus ou moins grande des observations est proche de son mode.

En effet, plus une forte proportion des individus prendra une valeur proche de celle du mode de la distribution, plus l'aplatissement sera faible.

On mesure l'aplatissement d'une courbe par comparaison à la courbe « normale » dont le moment centré d'ordre 4 est égal à 3.

Le degré d'aplatissement d'une courbe est caractérisé par la valeur du **coefficient d'aplatissement de Pearson** défini par :

$$K = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

On doit distinguer trois cas.

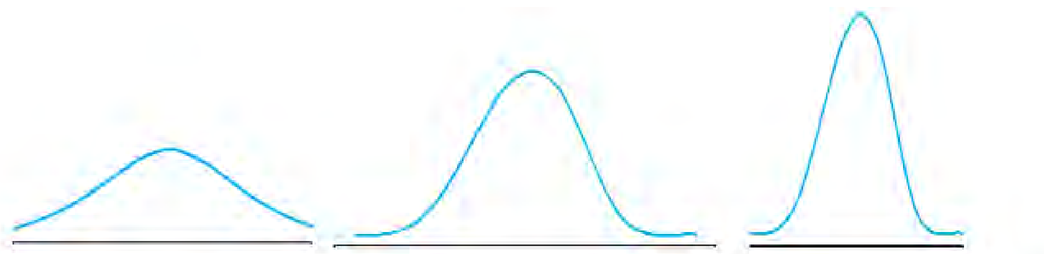
Cas 1 : si K < 3, la distribution est plus aplatie que la distribution normale.

On dit que la distribution est hyponormale ou platykurtique.

Cas 2 : si K = 3, la distribution est normale ou mésokurtique.

Cas 3 : si K > 3, la distribution est moins aplatie que la distribution

normale. On dit que la distribution est hypernormale ou leptokurtique.



Remarque :

Le coefficient d'asymétrie de Fisher et le coefficient d'aplatissement de Pearson sont sans dimension et indépendants d'un changement d'échelle et d'origine.

Exercice 13 : Calcul des caractéristiques de forme

Le tableau ci-dessous représente la distribution des notes obtenues à un devoir de Statistique pour 50 étudiants d'une faculté de médecine.

Notes	Effectifs
[0, 5[	4
[5, 10[	17
[10, 15[	26
[15, 20[	3
Total	**50**

1. Calculer les moments centrés d'ordre 2, 3 et 4.

2. Calculer et interpréter le coefficient d'asymétrie de Fisher et le coefficient d'aplatissement de Pearson.

Solution :

1. On dresse le tableau suivant, où x_i et n_i sont le centre et l'effectif de la classe numéro i, respectivement.

Notes	x_i	n_i	$n_i x_i$	$n_i(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^3$	$n_i(x_i - \bar{x})^4$
[0, 5[	2,5	4	10	243,36	–1898,208	14806,0224
[5, 10[	7,5	17	127,5	133,28	–373,184	1044,9152
[10, 15[	12,5	26	325	125,84	276,848	609,0656
[15, 20[	17,5	3	52,5	155,52	1119,744	8062,1568
Total		**50**	**515**	**658**	**–874,8**	**24522,16**

1.1 Moyenne arithmétique :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{4} n_i x_i = \frac{515}{50} = 10{,}3$$

1.2 Moment centré d'ordre deux ou variance

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{4} n_i (x_i - \bar{x})^2 = \frac{658}{50} = 13{,}2$$

1.3 Moment centré d'ordre trois

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{4} n_i (x_i - \bar{x})^3 = \frac{-874{,}8}{50} = -17{,}5$$

1.4 Moment centré d'ordre quatre

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{4} n_i (x_i - \bar{x})^4 = \frac{24522{,}16}{50} = 490{,}4$$

2. Calcul des caractéristiques de forme

2.1 Coefficient d'asymétrie de Fisher

$$S = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-17{,}5}{(13{,}2)^{1{,}5}} = -0{,}365$$

Comme S < 0, la distribution des notes est étalée vers la gauche.
La distribution des notes a un biais négatif.

2.2 Coefficient d'aplatissement de Pearson

$$K = \frac{\mu_4}{\mu_2^2} = \frac{490{,}4}{(13{,}2)^2} = 2{,}815$$

Comme K < 3, la distribution des notes est hyponormale ou platikurtique.

4. Caractéristiques de concentration

« Un homme en vaut cent et cent n'en valent pas un. »
MOULUC.

4.1 Introduction

4.1.1 Domaine d'application

Les indicateurs de dispersion peuvent parfois se révéler insuffisants pour traduire certains phénomènes. La notion de concentration est une notion connexe de celle de dispersion.

Les revenus des ménages d'un pays donné sont-ils concentrés ?
Telle branche industrielle est-elle concentrée ?

Une mesure de la concentration se révèle très utile pour répondre à de telles questions.

Deux conditions sont nécessaires pour étudier la concentration.

L'addition des différentes modalités du caractère doit avoir un sens ; additionner les tailles des individus d'une population n'aurait guère de signification.

Le partage de la masse globale du caractère doit être possible : partage des revenus entre les individus ; mais répartir la somme des tailles d'un groupe d'individus entre ces individus serait absurde.

La mesure des inégalités de revenu, au niveau mondial, au niveau national, avant ou après impôt, ou même au niveau d'une profession, constitue des exemples d'application de la notion de concentration.

4.1.2 Notion de valeur globale

On appelle valeur globale (ou masse de valeurs) d'une série statistique, les valeurs nixi, ni étant l'effectif de la classe de centre x_i.
On appelle valeur globale relative la quantité :

$$q_i = \frac{n_i x_i}{\sum_{i=1}^{p} n_i x_i}$$

On appelle valeur globale relative cumulée la quantité :

$$F_i = \sum_{j=1}^{i} q_j$$

4.1.3 Notion de concentration

Une série statistique est dite concentrée si une faible proportion des individus observés possède une forte proportion des valeurs globales.

Exemple 1 :

Soit la série statistique représentant les dépenses de santé.

Si 70% des dépenses de santé effectuées sont le fait de 10% des individus alors la concentration est forte.

La distribution les dépenses de santé est dite inégalitaire.

Exemple 2 :

On considère une autre série statistique représentant les dépenses de santé.
Si 33% des dépenses de santé effectuées sont le fait de 31% des individus alors la concentration est faible.
La distribution les dépenses médicales est dite égalitaire.

Les indicateurs de concentration les plus utilisés sont la courbe de concentration de Gini et l'indice de concentration de Gini.

4.2 L'indice de concentration de Gini

4.2.1 La courbe de concentration

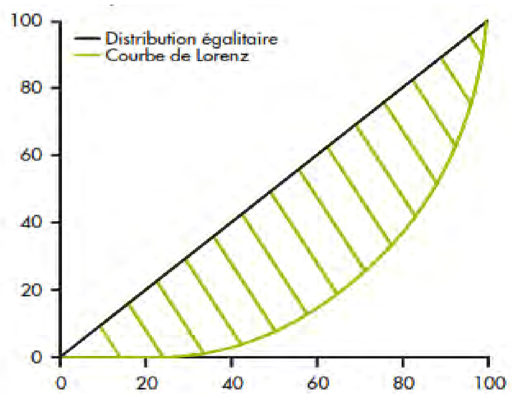
Cette courbe a été introduite par Gini en 1912, lors d'études sur les salaires et les revenus. Elle a pour but de décrire les effets de la concentration dans une population.

La courbe de concentration nécessite la détermination des fréquences relatives cumulées croissantes F_i et des valeurs globales relatives cumulées croissantes Q_i.
On a donc :

$$\forall\ i = 1,2,\cdots,p \quad \begin{cases} 0 \leq F_i \leq 1 \\ 0 \leq Q_i \leq 1 \end{cases}$$

On construit alors un carré de coté 1 sur lequel on porte en abscisses les F_i et en ordonnée les Q_i.

Ainsi on détermine (p+1) points (avec l'origine) à l'intérieur du carré ; la courbe qui joint ces (p+1) points est appelée courbe de concentration ou courbe de Gini ou courbe de Lorenz.



La courbe de Gini est située en dessous de la diagonale de référence.

Plus la courbe est éloignée de la diagonale, plus la concentration est forte.

4.2.2 L'indice de concentration de Gini

La courbe de Lorenz permet de calculer l'indice de Gini qui est une mesure du degré d'inégalité.

L'indice de Gini G correspond au rapport entre la surface hachurée et le triangle en dessous, il varie entre 0 et 1.
G est d'autant plus grand que la série considérée est concentrée.

Si l'indice est de 0, cela signifie que la courbe de Gini est la diagonale, l'égalité est parfaite.

Si l'indice est de 1, cela signifie qu'une seule personne détient tout le revenu, c'est l'inégalité maximale.

Plus la courbe est éloignée de la diagonale, plus l'indice de Gini est grand.

4.2.3 Calcul de l'indice de concentration avec la méthode des triangles

$$G = \sum_{j=1}^{p-1} (F_j Q_{j+1} - F_{j+1} Q_j)$$

où p est le nombre de classes.

Exercice 14 : Calcul des caractéristiques de concentration

On étudie 130 clients selon leur achat mensuel (en milliers de FCFA).

Montant des achats	Nombre de clients
[3 ; 10 [	20
[10, 50 [	50
[50, 100 [	18
[100, 500 [	22
[500, 1 000 [	16
[1 000, 3 000 [	4
Total	**130**

1. Quelle est la population étudiée ? Quelle est l'unité statistique ? Quel est l'échantillon observé ?

2. Quel est le caractère observé ? Quelle est sa nature ?

3. Tracer la courbe de concentration de Gini. Commenter.

4. Calculer l'indice de concentration de Gini par la méthode des triangles. Commenter.

5. Quelle est l'influence sur l'indice de Gini d'une augmentation de 4% des montants des achats ?

Solution :

1. La population étudiée est l'ensemble des clients. L'échantillon observé est 130 clients. L'unité statistique est un client.

2. Le caractère observé est le montant des achats. Sa nature est quantitative continue.

3. Courbe de concentration

Nous dressons le tableau suivant, où x_i et n_i sont le centre et l'effectif de la classe numéro i, respectivement.

Classes	n_i	x_i	f_i	F_i	n_ix_i	q_i	Q_i
[3 ; 10 [	20	6,5	0,154	0,154	130	0,004	0,004
[10, 50 [	50	30	0,385	0,538	1500	0,051	0,055
[50, 100 [	18	75	0,138	0,677	1350	0,046	0,101
[100, 500 [	22	300	0,169	0,846	6600	0,223	0,324
[500, 1 000 [	16	750	0,123	0,969	12000	0,406	0,73
[1 000, 3 000 [	4	2000	0,031	1	8000	0,27	1
Total	**130**		**1**		**29 580**	**1**	

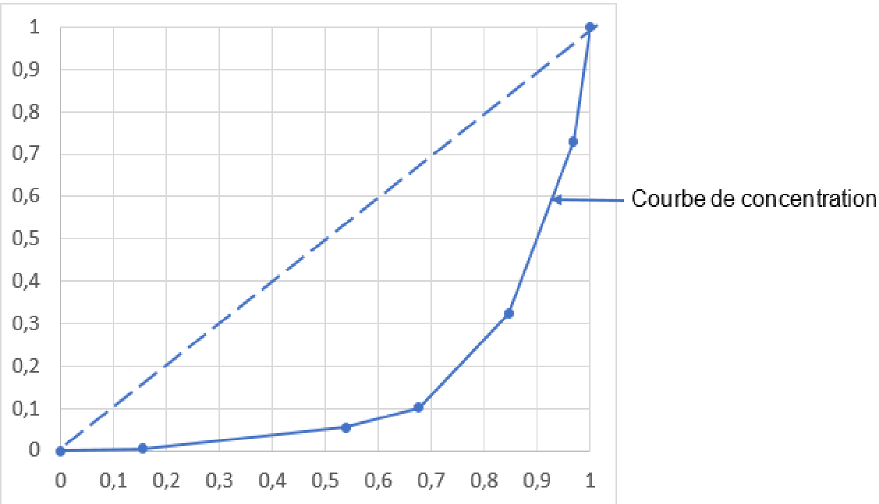
$$f_i = \frac{n_i}{\sum_i n_i} \quad ; \quad F_i = \sum_{j=1}^{i} f_j$$

$$q_i = \frac{n_i x_i}{\sum_i n_i x_i} \quad ; \quad Q_i = \sum_{j=1}^{i} q_j$$

f_i : fréquence relative ; F_i : fréquence relative cumulée croissante
q_i : valeur globale relative ; Q_i : valeur globale relative cumulée croissante

On construit alors un carré de coté 1 sur lequel on porte en abscisses les fréquences relatives cumulées croissantes F_i et en ordonnée les valeurs globales relatives cumulées croissantes Q_i.

F_i	0,154	0,538	0,677	0,846	0,969	1
Q_i	0,004	0,055	0,101	0,324	0,73	1



L'aire de concentration est la surface comprise entre la diagonale de référence (droite en pointillés bleus) et la courbe de concentration (courbe bleue).
Comme l'aire de concentration est grande. La distribution des montants des achats est fortement concentrée.

4. Calcul de l'indice de Gini par la méthode des triangles

i	F_i	Q_i	Gini = $F_i\ Q_{i+1} - F_{i+1}\ Q_i$
1	0,154	0,004	0,006
2	0,538	0,055	0,017
3	0,677	0,101	0,134
4	0,846	0,324	0,303
5	0,969	0,730	0,240
6	1	1	
			Indice de Gini = 0,7

$$i = 1 \qquad F_1Q_2 - F_2Q_1 = (0{,}154 \times 0{,}055) - (0{,}538 \times 0{,}004) = 0{,}006$$

$$i = 2 \qquad F_2Q_3 - F_3Q_2 = (0{,}538 \times 0{,}101) - (0{,}677 \times 0{,}055) = 0{,}017$$

$$i = 3 \qquad F_3Q_4 - F_4Q_3 = (0{,}677 \times 0{,}324) - (0{,}846 \times 0{,}101) = 0{,}134$$

$$i = 4 \qquad F_4Q_5 - F_5Q_4 = (0{,}846 \times 0{,}73) - (0{,}969 \times 0{,}324) = 0{,}303$$

$$i = 5 \qquad F_5Q_6 - F_6Q_5 = (0{,}969 \times 1) - (1 \times 0{,}73) = 0{,}24$$

G = 0,006 + 0,017 + 0,134 + 0,303 + 0,24 = 0,7

L'indice de concentration de Gini est égal à 0,7.

Comme la valeur de l'indice de Gini est supérieure à 0,35 ; la distribution des montants des achats est fortement concentrée.
La distribution des montants des achats est inégalitaire.

5. On a :

$$q_i = \frac{n_i x_i}{\sum_{i=1}^{p} n_i x_i}$$

Si x_i augmente de 4%, on a alors

$$q_i = \frac{n_i (x_i + 0{,}04\ x_i)}{\sum_{i=1}^{p} n_i (x_i + 0{,}04\ x_i)} = \frac{n_i x_i}{\sum_{i=1}^{p} n_i x_i}$$

Le nombre q_i reste inchangé. Une augmentation de 4% des montants des achats n'a aucune influence sur l'indice de concentration de Gini.

Exercice 15 : Calcul des caractéristiques de tendance centrale, de dispersion, de forme et de concentration

Logiciel : Stata

Dans une entreprise les 200 factures établies dans le courant d'un certain mois ont été classées d'après leur montant.

Les résultats de l'opération sont consignés dans le tableau suivant :

3	7	4	3	7	4	8	7	102	107
7	5	7	7	5	7	8	5	454	396
8	8	6	8	8	6	7	8	409	232
3	6	5	3	6	5	7	6	365	321
7	8	8	7	8	8	6	8	232	456
5	7	8	5	7	8	4	7	150	176
6	5	6	6	5	6	4	5	498	198
2	3	6	2	3	6	3	3	187	498
9	7	7	9	7	7	5	7	176	432
7	6	8	7	6	8	6	6	232	287
12	17	11	12	17	11	1754	54	505	515
11	49	17	11	49	17	2356	56	597	765
43	47	19	43	47	19	1287	76	758	675
32	23	21	32	23	21	1500	86	876	786
42	17	23	42	17	23	53	90	543	897
17	15	25	17	15	25	67	94	876	786
19	12	29	19	12	29	78	65	953	76
47	24	24	47	24	24	75	57	654	87
43	43	34	43	43	34	84	95	670	97
23	41	45	23	41	45	57	76	876	83

Les montants des factures sont exprimés en milliers de FCFA.

1. Quelle est la population étudiée ? Quel est l'échantillon observé ?
Quelle est l'unité statistique ?

2. Quel est le caractère observé ? Quelle est sa nature ?

3. Tracer l'histogramme de cette distribution.

4. Calculer les caractéristiques de tendance centrale suivantes : moyenne arithmétique, quantiles (3 quartiles, premier centile, quatre-vingt-dix neuvième centile, premier décile, neuvième décile) et mode.

5. Calculer les caractéristiques de dispersion suivantes : étendue, intervalle interquartile, intervalle interdécile, intervalle intercentile, variance, écart-type et coefficient de variation.

6. Calculer les caractéristiques de forme suivantes : coefficient de dissymétrie de Fisher et coefficient d'aplatissement de Pearson.

7. Tracer la courbe de concentration avec le logiciel Stata.

8. Calculer l'indice de Gini avec le logiciel Stata.

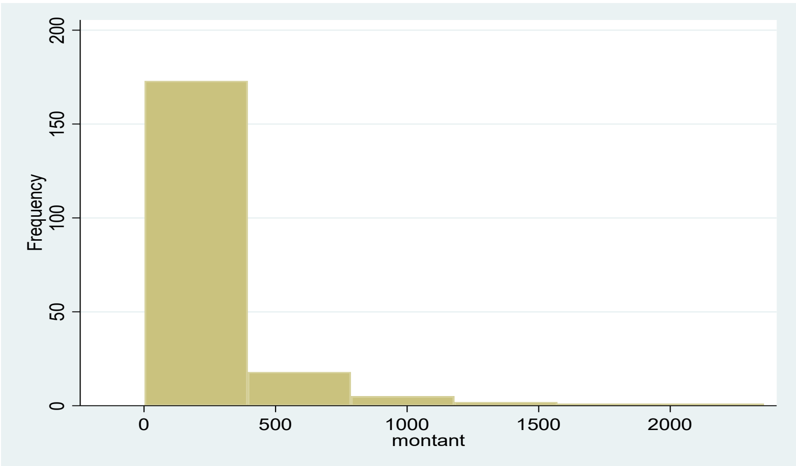
Solution

1.
1.1 La population étudiée est l'ensemble des factures.
1.2 L'échantillon observé est les 200 factures.
1.3 L'unité statistique est une facture.

2.
2.1 Le caractère observé est le montant des factures.
2.2 Le montant des factures est un caractère quantitatif continu.

3. Histogramme avec Stata

```
histogram   montant,  freq   bin(6)
```



La commande

```
sum  montant ,   detail
```

donne les statistiques descriptives de la série « montant ».

```
                        montant
-------------------------------------------------------------
      Percentiles      Smallest
 1%          2.5              2
 5%          3.5              2
10%            5              3       Obs                 200
25%            7              3       Sum of Wgt.         200

50%           19                      Mean            140.905
                        Largest       Std. Dev.      314.9014
75%           76           1287
90%        501.5           1500       Variance       99162.86
95%          786           1754       Skewness       3.693735
99%         1627           2356       Kurtosis       19.88152
```

4. Caractéristiques de tendance centrale

4.1 La moyenne arithmétique **(Mean)**

Mean : 140.905

Le montant moyen des factures est égal à 140 905 FCFA.

4.2 Les quantiles **(Percentiles)**

4.2.1 Les quartiles

a. Le premier quartile est le quantile d'ordre 25%.
Q_1 = 7 → 25% des factures ont un montant inférieur à 7 000 F CFA.

b. Le deuxième quartile (ou la médiane) est le quantile d'ordre 50%.
Q_2 = 19 → 50% des factures ont un montant inférieur à 19 000 FCFA.
c. Le troisième quartile est le quantile d'ordre 75%.
Q_3 = 76 → 75% des factures ont un montant inférieur à 76 000 FCFA.

4.2.2 Autres quantiles

a. Le premier centile est le quantile d'ordre 1%.
C_1 = 2,5 → 1% des factures ont un montant inférieur à 2 500 FCFA.

b. Le quatre-vingt-dix neuvième centile est le quantile d'ordre 99%.
C_{99} = 1 627 → 99% des factures ont un montant inférieur à 1 627 000 FCFA.

c. Le neuvième décile est le quantile d'ordre 90%.
D_9 = 501,5 → 90% des factures ont un montant inférieur à 501 500 FCFA.

d. Le premier décile est le quantile d'ordre 10%.
D_1 = 5 → 10% des factures ont un montant inférieur à 5 000 FCFA.

4.3 Le mode

Le mode Mo d'une distribution statistique est sa valeur la plus fréquente.

Le montant des factures est un caractère quantitatif continu, son mode est indéterminé.

Pour un caractère quantitatif continu, on peut seulement déterminer la classe modale.

4.3.1 Détermination de la classe modale

Comme les classes n'ont pas la même amplitude, la classe modale est celle de densité maximale.

La densité d'une classe est le rapport de l'effectif à l'amplitude.

$$d_i = \frac{n_i}{a_i}$$

Montant	n_i	a_i	d_i	f_i	F_i
[0 ; 10[	80	10	8	0,4	0,4
[10 ; 50[	60	40	1,5	0,3	0,7
[50 ; 100[	20	50	0,4	0,1	0,8
[100 ; 500[	20	400	0,05	0,1	0,9
[500 ; 1000[	16	500	0,032	0,08	0,98
[1000 ; 3000[	4	2000	0,002	0,02	1
Total	200			1	

La classe modale est [0,10[, cette classe a pour densité 8.
Les montants des factures les plus fréquents sont compris entre 0 et 10 000 FCFA.

4.3.2 Calcul du mode avec Stata

Le mode est indéterminé, mais le logiciel Stata calcule le mode.

```
egen  modemt =  mode(montant) , minmode
```

```
list  modemt  in 1
```

```
     +--------+
     | modemt |
     |--------|
  1. |      7 |
     +--------+
```

Le mode est égal à 7 000 FCFA, le montant des factures le plus fréquent est 7 000 FCFA.

5. Caractéristiques de dispersion

5.1 Etendue

L'étendue W est égale à la différence entre la valeur maximum observée et la valeur minimum observée.

Le montant minimal est 2 000.
Le montant maximal est 2 356 000.

L'étendue vaut

$$W = X_{max} - X_{min} = 2\,356\,000 - 2\,000 = 2\,354\,000$$

Calcul de l'étendue avec Stata

```
tabstat  montant , stat(range)
```

```
    variable |     range
-------------+----------
     montant |      2354
------------------------
```

5.2 Intervalles interquantiles

5.2.1 Intervalle interquartile

$[Q_1 , Q_3[$ est l'intervalle interquartile, il contient 50% des observations.

L'intervalle interquartile est [5 ; 76[, 50% des factures ont un montant compris entre 5 000 et 76 000 FCFA.

5.2.2 Intervalle interdécile

$[D_1 , D_9[$ est l'intervalle interdécile, il contient 80% des observations.
L'intervalle interdécile est [5 ; 501,5[, 80% des factures ont un montant compris entre 5 000 et 501 500 FCFA.

5.2.3 Intervalle intercentile

$[C_1 , C_{99}[$ est l'intervalle interdécile, il contient 98% des observations.
L'intervalle interdécile est [2,5 ; 1627[, 98% des factures ont un montant entre 2500 et 1 627 000 FCFA.

5.3 La variance **(variance)**

La variance d'une variable X est son moment centré d'ordre 2 :

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

Le logiciel Stata calcule la variance corrigée.

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

La variance du montant des factures est égale à 99 162,86.

La variance est en (milliers de FCFA)².

5.4 L'écart type **(standard deviation)**

L'écart type d'une série est égal à la racine carrée de la variance :

$$\sigma(X) = \sqrt{Var(X)}$$

L'écart type du montant des factures est égal à 314 901,4 FCFA

La dispersion du montant des factures autour du montant moyen est égale à 314 901,4 FCFA.

L'écart type est exprimé dans la même unité que les observations de la série, ce qui en statistique descriptive, rend le résultat plus facilement interprétable.

5.5 Le coefficient de variation

Le coefficient de variation est une mesure de dispersion relative. Il est le rapport de l'écart type σ à la moyenne $\bar{x}$:

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

Le coefficient de variation est égal à :

$$CV = \frac{\sigma}{\bar{x}} = \frac{314,9014}{140,905} = 2,235$$

soit 223,5%.

Le coefficient de variation est supérieur à 30%. La distribution des montants des factures est fortement dispersée et les observations sont hétérogènes.
Calcul du coefficient de variation avec Stata

```
tabstat  montant , stat(cv)
```

```
    variable  |        cv
-------------+----------
     montant |   2.234849
------------------------------------
```

Remarque :

Comme le coefficient de variation dépasse 30%, alors les observations sont dispersées et il devient hasardeux d'interpréter la moyenne.

6. Caractéristiques de forme

Les principales caractéristiques en ce qui concerne la forme sont : la dissymétrie et l'aplatissement.

6.1 Le coefficient d'asymétrie de Fisher **(Skewness)**

Le coefficient d'asymétrie de Fisher est égal à 3,69.
Comme cette valeur est positive, la distribution des montants des factures est étalée vers la droite.
La distribution des montants a un biais positif.

6.2 Le coefficient d'aplatissement de Pearson **(Kurtosis)**

Le coefficient d'aplatissement de Pearson est égal à 19,88.

Comme cette valeur est supérieure à 3, la distribution des montants des factures est hypernormale.

La distribution des montants des factures est leptokurtique.

7. Tracé de la courbe de Gini avec Stata

Attention, il faut télécharger à partir d'internet le package lorenz.

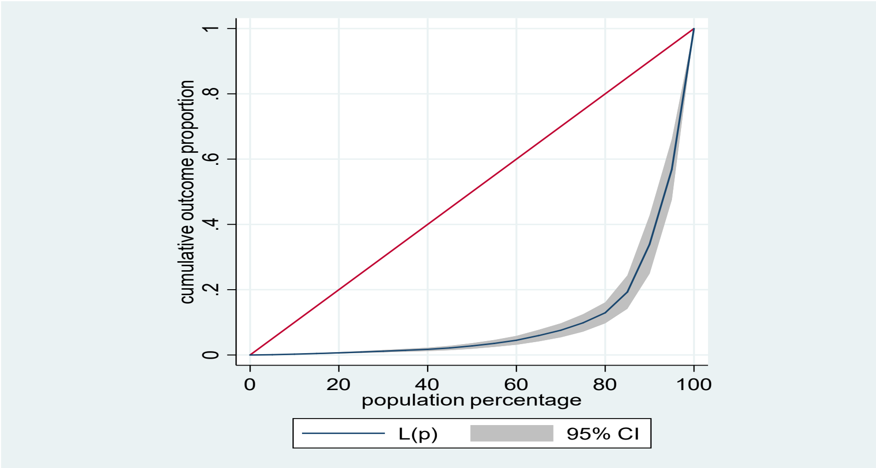
Saisir dans le fenêtre « Command », l'instruction suivante :

```
ssc  install  lorenz,  all
```

Valider par la touche « Enter » du clavier.

```
lorenz  estimate     montant
```

```
lorenz  graph , aspectratio(1) xlabel (, grid)
```



8. Calcul de l'indice de Gini avec Stata

Attention, il faut télécharger à partir d'internet le package ineqerr.
Saisir dans le fenêtre « Command », l'instruction suivante :

```
ssc install ineqerr, all
```

Valider par la touche « Enter » du clavier.

```
ineqerr   montant, reps(50)
```

```
montant ---------------------------------------------------------- (unlabeled)
(obs=200)

Bootstrap statistics

Variable |    Reps    Observed       Bias    Std. Err.    [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Gini |      50      .79069  -.0041741    .0142531     .7620472   .8193328   (N)
         |                                                 .759095   .8136063   (P)
         |                                                .7725182   .8145092  (BC)
---------+--------------------------------------------------------------------
   Theil |      50    1.313892  -.0104744    .0874223      1.13821   1.489573   (N)
         |                                                1.138535   1.487601   (P)
         |                                                1.138535   1.507533  (BC)
---------+--------------------------------------------------------------------
 Varlogs |      50    2.918244  -.0503633    .2288244     2.458404   3.378084   (N)
         |                                                2.534501     3.3328   (P)
         |                                                2.550129   3.380003  (BC)
------------------------------------------------------------------------------
                              N = normal, P = percentile, BC = bias-corrected
```

Le coefficient de Gini est égal à 0,79.

Comme la valeur de G est proche de 1, la concentration est forte.

On a une forte concentration du chiffre d'affaires sur un petit nombre de factures.

La distribution des montants des factures est inégalitaire.

Objectifs pédagogiques

A la fin de l'étude du chapitre 3, l'auditeur sera capable de :

1. définir la notion d'indice élémentaire ;
2. établir les deux propriétés des indices élémentaires : réversibilité et circularité ;
3. définir la notion d'indice synthétique ;
4. utiliser les quatre formules d'indices synthétiques : valeur, Laspeyres, Paasche, Fisher ;
5. établir les relations entre les différents indices ;
6. définir et utiliser les indices chaînes.

1. Introduction

Les indices ont pris ces dernières années, une place importante dans l'étude de l'activité économique et financière : indices des prix, de production, du commerce extérieur ; indice boursier, etc.

Leur succès provient de leur qualité principale, à savoir qu'ils permettent de formuler en un seul nombre la synthèse de nombreuses mesures. En outre, ils sont un instrument des études chronologiques : évolution des prix à la consommation, etc.

De façon générale, un indice sert à mesurer une variation relative entre deux situations d'une grandeur que l'on dira simple ou complexe.

Une grandeur simple est une grandeur entièrement définie par la donnée d'un seul nombre.

Le prix du ticket de bus, le taux de fécondité, la production d'acier sont des exemples de grandeurs simples.

Une grandeur complexe est une grandeur caractérisée par la donnée de plusieurs nombres. Une grandeur complexe est donc l'énoncé de plusieurs grandeurs simples.

Le niveau général des prix, la production industrielle, les exportations sont des exemples de grandeurs complexes.

Le niveau général des prix est constitué des prix des divers aliments et boissons, du logement, de l'équipement ménager, de l'habillement, des services médicaux, des transports, etc.
Dans l'étude des phénomènes économiques et sociaux, on a souvent besoin de décrire les variations de grandeurs simples ou complexes. Ces comparaisons dans le temps et dans l'espace, se font généralement en effectuant le rapport des grandeurs considérées.

Dans le cas où l'on effectue le rapport de grandeurs simples : on parle d'indice statistique élémentaire.

Dans le cas où l'on effectue le rapport de grandeurs complexes : on parle d'indice synthétique.

2. Les indices élémentaires

Dans ce qui suit, nous traiterons les variations entre deux dates, les concepts étant facilement transposables aux variations entre deux lieux.

2.1 Définition

Considérons l'évolution temporelle d'une grandeur simple X : $x_0, x_1, x_2, \dots, x_t, \dots$

$x_0, x_1, x_2, \dots, x_t$ sont donc les valeurs prises par la grandeur X, aux dates successives 0, 1, 2,…, t, …

On appelle indice élémentaire de la grandeur simple X à la date t par rapport à la date 0, le rapport :

$$i_{t/o}(X) = \frac{x_t}{x_o}$$

La date (ou période) 0 est la date de référence ou de base de l'indice, cette date est utilisée comme date de comparaison.

La date (ou période) t est la date courante.

Généralement, ce rapport est exprimé en pourcentage :

$$I_{t/o} = 100 \times i_{t/o} = 100 \times \frac{x_t}{x_o}$$

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

Un indice élémentaire est donc le rapport de deux valeurs prises par une même grandeur simple à deux dates différentes.

C'est un nombre sans dimension, indépendant des unités choisies pour mesurer la grandeur statistique.

Exercice 1

1. Le nombre de malades internés dans un hôpital est passé de 262 806 en 2018 à 305 666 en 2019.

L'indice du nombre de malades internés en 2019 par rapport à 2018 est :

$$I_{2019/2018}(\text{Nmal}) = 100 \times \frac{305\,666}{262\,806} = 116{,}31$$

soit une augmentation de 16,31%.

2. Le prix d'un produit de grande consommation est passé de 1 020 FCFA en 2018 à 950 FCFA en 2019.

L'indice du prix du produit de 2019 par rapport à 2018 est :

$$I_{2019/2018}(\text{Prix}) = 100 \times \frac{950}{1020} = 93{,}14$$

soit une baisse de 6,86%.

2.2 Propriétés des indices élémentaires

Les indices élémentaires possèdent les deux propriétés de réversibilité et de circularité (transitivité).

2.2.1 Réversibilité

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

En effet :

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

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

2.2.2 Circularité

$$i_{t/o} = i_{t/t*} \, . \, i_{t*/o}$$

En effet :

$$\frac{x_t}{x_o} = \frac{x_t}{x_{t*}} . \frac{x_{t*}}{x_o}$$

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

$$i_{t/t^*} = \frac{i_{t/o}}{i_{t^*/o}}$$

Exercice 2 : Prix d'un médicament

2015 : 404,90 FCFA
2017 : 494,58 FCFA
2019 : 687,19 FCFA

Les indices élémentaires du prix du médicament, base 100 en 2015 sont :

$$I_{2017/2015} = 100 \times \frac{494{,}58}{404{,}90} = 122{,}15$$

$$I_{2019/2015} = 100 \times \frac{687{,}19}{404{,}90} = 169{,}72$$

$$I_{2019/2017} = 100 \times \frac{687{,}19}{494{,}58} = 138{,}94$$

Mais on peut aussi calculer l'indice $I_{2019/2017}$ en ignorant les prix aux différentes périodes :

$$I_{2019/2017} = 100 \times \frac{i_{2019/2015}}{i_{2017/2015}} = 100 \times \frac{1{,}6972}{1{,}2215} = 138{,}94$$

$$I_{2015/2019} = 100 \times \frac{1}{i_{2019/2015}} = 100 \times \frac{1}{1{,}6972} = 58{,}92$$

On a utilisé la propriété de réversibilité.

2.3 Pourcentage de variation

On appelle pourcentage de variation le nombre :

$$PV_{t/o} = \frac{x_t - x_o}{x_o} \times 100$$

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

$$= I_{t/o} - 100$$

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

i. Si PV > 0 alors on a une augmentation de la variable.

ii. Si PV = 0 alors on a une stagnation de la variable.

iii. Si PV < 0 alors on a une diminution de la variable.

Exercice 3

Le prix d'un médicament est passé de 404,9 FCFA en 2019 à 687,19 FCFA en 2020.

Le pourcentage de variation est :

$$PV_{2020/2019} = \left(\frac{687,19 - 404,90}{404,90}\right) \times 100$$

$$= 100 \times (i_{2020/2019} - 1) = 100 \times (1,6972 - 1) = 69,72\%$$

Ce qui correspond bien à l'indice $I_{2020/2019}$ = 169,72.

Le prix du médicament a augmenté de 69,72%.

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

Exercice 4 : Prix d'un bien quelconque

Novembre 2018 : 200 FCFA

Novembre 2019 : 250 FCFA

Novembre 2020 ; 400 FCFA

1. Augmentation de novembre 2018 à novembre 2019

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

2. Augmentation de novembre 2019 à novembre 2020

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

3. Augmentation de novembre 2018 à novembre 2020

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

et non pas 60 + 25 = 85%.

Exercice 5 : Calcul du taux d'intérêt réel

On emprunte pour un an un capital au taux d'intérêt 11%. Pendant ce temps, les prix augmentent de 5%.
Déterminer le taux d'intérêt réel de cet emprunt.

Solution

Un capital K_0, emprunté au taux d'intérêt a, doit être remboursé à l'issue de la période :
$K_1 = K_0 (1 + a)$, intérêt compris,

Le rapport :

$$\frac{K_1}{K_0} = 1 + a$$

mesure l'évolution de la dette, accrue par l'intégration de l'intérêt.

Mais, en fait, en période d'inflation de taux b, les sommes K_0 et K_1 ne sont pas comparables.

Il faut donc convertir le capital emprunté K_0 en francs courants de la période ; soit K'_0 la valeur convertie :

$$K'_0 = K_0(1 + b)$$

Le rapport des sommes K_1 et K'_0 représentent l'augmentation en termes réels de la dette sur la période.

On a :

$$\frac{K_1}{K'_0} = \frac{K_0(1+a)}{K_0(1+b)} = \frac{1+a}{1+b}$$

L'écart :

$$\frac{K_1}{K'_0} - 1 = \frac{1+a}{1+b} - 1$$

est le taux d'intérêt réel.

Le calcul donne :

$$\frac{K_1}{K'_0} - 1 = \frac{1{,}11}{1{,}05} - 1 = 1{,}057 - 1 = 0{,}057$$

Le taux d'intérêt réel est de 5,7%.

Nous vérifions que contrairement à une opinion couramment reçue, le taux d'intérêt réel n'est pas (rigoureusement) égal au taux nominal déduction faite du taux d'inflation (taux de variation des prix).

3. Les indices synthétiques

Nous avons vu que les indices élémentaires retracent l'évolution d'une seule grandeur. Mais le plus souvent, l'économiste ou le manager, désire suivre les variations de grandeurs complexes : volume des importations, indice général des prix, production industrielle, etc.

Ces grandeurs complexes sont composées d'un nombre plus ou moins grand de grandeurs simples : par exemple le niveau général des prix est constitué des prix des services médicaux, des divers aliments et boissons, du logement, de l'équipement ménager, de l'habillement, des transports, etc.

L'évolution de chacune de ces grandeurs simples est décrite par un indice élémentaire.

L'opération de construction d'un indice synthétique relatif à la variation d'une grandeur complexe consiste donc à résumer une série d'indices élémentaires.

Trois formules d'indices synthétiques sont utilisées en pratique : les formules de Laspeyres, de Paasche et de Fisher.

3.1 Coefficient budgétaire

Considérons n produits pj dont on connaît les prix et les quantités à deux périodes différentes.

On appelle coefficient budgétaire (coefficient de pondération) du produit p_j, par rapport à l'une de ces périodes, la part représentée par la valeur (prix par quantité) de ce produit à cette période relativement à la valeur totale de tous les produits à cette même période.

Notons $c_{j,0}$ le coefficient budgétaire par rapport à la période de base :

$$c_{j,0} = \frac{p_{j,0} \times q_{j,0}}{\sum_{j=1}^{n} (p_{j,0} \times q_{j,0})}$$

Notons $c_{j,t}$ le coefficient budgétaire par rapport à la période courante :

$$c_{j,t} = \frac{p_{j,t} \times q_{j,t}}{\sum_{j=1}^{n} (p_{j,t} \times q_{j,t})}$$

On a bien :

$$\sum_{j=1}^{n} c_{j,0} = \sum_{j=1}^{n} c_{j,t} = 1$$

La somme des coefficients budgétaires est toujours égale à un.

3.2 Les différents types d'indices synthétiques

En calculant un indice synthétique, on peut chercher à saisir des variations de valeur, de prix ou de quantité (volume).

Quel que soit le cas, le calcul de l'indice met toujours en jeu à la fois des prix et des quantités.
Soit le produit P_j.

$p_{j,0}$ et $q_{j,0}$ représentent les prix et les quantité du produit P_j à la période de base.

$p_{j,t}$ et $q_{j,t}$ représentent les prix et les quantité du produit P_j à la période courante.

3.2.1 Indice de valeur

Un indice de valeur est le rapport de la somme des valeurs (produit du prix par la quantité correspondante) relatives à la période courante à la somme des valeurs de la période de base :

$$V_{t/0} = \frac{\sum_{j=1}^{n} (p_{j\,t}\ q_{j,t})}{\sum_{j=1}^{n} (p_{j,0}\ q_{j,0})}$$

Par exemple l'indice de la valeur courante des importations est le rapport de la valeur totale des importations pendant la période courante à leur valeur totale pendant la période de base.

Un indice de valeur n'a pas une grande signification économique car son évolution dépend aussi bien des variations des prix des marchandises que des variations des quantités importées, facteurs qu'il est utile de distinguer.

3.2.2 Indice de Laspeyres

Laspeyres (1834 – 1913) : mathématicien et économiste allemand.

L'indice de Laspeyres est la moyenne arithmétique des indices élémentaires, pondérée par les coefficients $c_{j,0}$ de la période de base :

$$L_{t/0} = \sum_{j=1}^{n} c_{j,0} \frac{x_{j,t}}{x_{j,0}}$$

Dans l'indice de Laspeyres, les coefficients de pondération sont fixes, ce sont ceux de la période de base.

3.2.2.1 Indice de Laspeyres des prix

$$L(p)_{t/0} = \sum_{j=1}^{n} c_{j,0} \cdot \frac{p_{j,t}}{p_{j,0}}$$

$$= \frac{\sum_{j=1}^{n} p_{j,0}\, q_{j,0} \frac{p_{j,t}}{p_{j,0}}}{\sum_{j=1}^{n} p_{j,0}\, q_{j,0}} = \frac{\sum_{j=1}^{n} p_{j,t}\, q_{j,0}}{\sum_{j=1}^{n} p_{j,0}\, q_{j,0}}$$

Ainsi l'indice de Laspeyres des prix apparaît comme le rapport :

$$\frac{\sum_{j=1}^{n} p_t\, q_0}{\sum_{j=1}^{n} p_0\, q_0} = \frac{\text{Dépense totale de la période de base évaluée aux prix courants.}}{\text{Dépense totale de la période de base}}$$

Dans le rapport, les quantités sont constantes et les prix variables parce que c'est un indice de prix.

3.2.2.2 Indice de Laspeyres des quantités

$$L(q)_{t/0} = \sum_{j=1}^{n} c_{j,0} \frac{q_{j,t}}{q_{j,0}}$$

$$= \frac{\sum_{j=1}^{n} p_{j,0}\, q_{j,0} \frac{q_{j,t}}{q_{j,0}}}{\sum_{j=1}^{n} p_{j,0}\, q_{j,0}} = \frac{\sum_{j=1}^{n} p_{j,0}\, q_{j,t}}{\sum_{j=1}^{n} p_{j,0}\, q_{j,0}}$$

Ainsi l'indice de Laspeyres des quantités apparaît comme le rapport :

$$\frac{\sum_{j=1}^{n} p_0\, q_t}{\sum_{j=1}^{n} p_0\, q_0} = \frac{\text{Dépense totale de la période courante évaluée aux prix de la période de base.}}{\text{Dépense totale de la période de base}}$$

3.2.3 Indice de Paasche

Paasche (1851-1925) : statisticien allemand.

L'indice de Paasche est la moyenne harmonique des indices élémentaires pondérée par les coefficients $c_{j,t}$ de la période courante :

$$\frac{1}{P_{t/0}} = \sum_{j=1}^{n} c_{j,t}\ \frac{1}{\frac{x_{j,t}}{x_{j,0}}}$$

$$\frac{1}{P_{t/0}} = \sum_{j=1}^{n} c_{j,t}\ \frac{1}{i_{t/0}\ (X)}$$

Dans l'indice de Paasche les coefficients de pondération sont ceux de la période courante.

3.2.3.1 Indice de Paasche des prix

$$P(p)_{t/0} = \frac{1}{\sum_{j=1}^{n} c_{j,t}\ \frac{p_{j,0}}{p_{j,t}}}$$

$$= \frac{\sum_{j=1}^{n} p_{j,t}\ q_{j,t}}{\sum_{j=1}^{n} p_{j,t}\ q_{j,t}\ \times \frac{p_{j,0}}{p_{j,t}}} = \frac{\sum_{j=1}^{n} p_{j,t}\ q_{j,t}}{\sum_{j=1}^{n} p_{j,0}\ q_{j,t}}$$

Ainsi l'indice de Paasche des prix apparaît comme le rapport :

$$\frac{\sum_{j=1}^{n} p_t\ .q_t}{\sum_{j=1}^{n} p_0\ .q_t} = \frac{\text{Dépense totale de la période courante.}}{\text{Dépense totale de la période courante évaluée aux prix de l'année de base}}$$

3.2.3.2 Indice de Paasche des quantités

$$P(q)_{t/0} = \frac{1}{\sum_{j=1}^{n} c_{j,t} \times \frac{q_{j,0}}{q_{j,t}}}$$

$$= \frac{\sum_{j=1}^{n} p_{j,t}\, q_{j,t}}{\sum_{j=1}^{n} p_{j,t}\, q_{j,t} \times \frac{q_{j,0}}{q_{j,t}}} = \frac{\sum_{j=1}^{n} p_{j,t}\, q_{j,t}}{\sum_{j=1}^{n} p_{j,t}\, q_{j,0}}$$

Ainsi l'indice de Paasche des prix apparaît comme le rapport :

$$\frac{\sum_{j=1}^{n} p_t q_t}{\sum_{j=1}^{n} p_t q_0} = \frac{\text{Dépense totale de la période courante.}}{\text{Dépense totale de la période de base évaluée aux prix de l'année courante}}$$

3.2.4 Indice de Fisher

Irving Fisher (1867-1947) : statisticien américain.

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

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

3.2.4.1 Indice de Fisher des prix

$$F(p)_{t/0} = \sqrt{L(p)_{t/0} \,.\, P(p)_{t/0}}$$

3.2.4.2 Indice de Fisher des quantités

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

Remarque

L'indice de Laspeyres a tendance à surestimer une hausse des prix, tandis que l'indice de Paasche a tendance à la sous-estimer.

On en déduit que l'indice de Fisher doit donner une meilleure estimation d'une hausse des prix.

Nous savons aussi que la moyenne harmonique est inférieure à la moyenne arithmétique, ceci fait que l'indice de Paasche est souvent plus petit que l'indice de Laspeyres.

Cependant, la comparaison des indices de Laspeyres et de Paasche ne peut se faire que si les coefficients de pondération sont les mêmes.

Exercice 6

Calculer les indices de valeur, de Laspeyres, de Paasche et de Fisher pour Décembre 2018 par rapport à Novembre 2018 sur l'ensemble des quatre produits décrits ci-dessous :

Produits	Novembre 2018		Décembre 2018	
	Prix p_0	Quantités q_0	Prix p_t	Quantités q_t
A	9,00	27	9,25	37
B	4,90	31	5,20	40
C	3,65	40	5,00	28
D	8,10	15	7,70	30

Les prix sont donnés en centaines de FCFA et les quantités sont des nombres.

Solution

Le mois de novembre 2018 correspond à la période 0.

Le mois de décembre 2018 correspond à la période t.

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

	j	p_0q_0	p_0q_t	p_tq_0	p_tq_t
A	1	243,00	333,00	249,75	342,25
B	2	151,90	196,00	161,20	208,00
C	3	146,00	102,20	200,00	140,00
D	4	121,50	243,00	115,50	231,00
	Total	**662,40**	**874,20**	**726,45**	**921,25**

1. Indice de valeur $V_{t/0}$

$$V_{t/0} = \frac{\sum p_t q_t}{\sum p_0 q_0} \times 100 = \frac{921,25}{662,40} \times 100 = 139,07$$

Soit une augmentation de 39,07% de la valeur.

2. Indice de Laspeyres

2.1 Indice de Laspeyres des prix

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

Soit une augmentation de 9,67% des prix.

2.2 Indice de Laspeyres des quantités

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

Soit une augmentation de 31,97% des quantités.

3. Indice de Paasche

3.1 Indice de Paasche des prix

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

Soit une augmentation de 5,38% des prix.

3.2 Indice de Paasche des quantités

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

Soit une augmentation de 26,81% des quantités.

4. Indice de Fisher

4.1 Indice de Fisher des prix

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

Soit une augmentation de 7,5% des prix.

4.2 Indice de Fisher des quantités

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

Soit une augmentation de 29,36% des quantités.

3.2.5 Utilisation des indices

L'indice de Laspeyres est le plus commode à utiliser ; la plupart des indices courants établis par les instituts du monde entier sont du type « Laspeyres ».

L'indice de Paasche, symétrique de celui de Laspeyres quant à sa signification, présente des inconvénients pratiques à cause de la mise à jour permanente de ses pondérations.

Il n'est, de ce fait, pas utilisé dans le calcul direct des indices courants. Son calcul est néanmoins intéressant pour obtenir, avec l'indice de Laspeyres, une fourchette d'estimation.

L'indice de Fisher est quasiment inusité car son calcul ne peut pas se faire par agrégation progressive.

3.3 Propriétés des indices de Laspeyres, de Paasche et de Fisher

3.3.1 Réversibilité

3.3.1.1 Indice de Laspeyres

$$L_{0/t} = \sum_j c_{j,t} \frac{x_{j,0}}{x_{j,t}} = \frac{1}{P_{t/0}}$$

car

$$\frac{1}{P_{t/0}} = \sum_j c_{j,t} \frac{1}{i_{t/0}(X)}$$

On a :

$$L_{0/t} \neq \frac{1}{L_{t/0}}$$

L'indice de Laspeyres n'est pas réversible.

3.3.1.2 Indice de Paasche

$$\frac{1}{P_{0/t}} = \sum_j c_{j,0} \frac{x_{j,t}}{x_{j,0}}$$

$$P_{0/t} = \frac{1}{\sum_j c_{j,0} \frac{x_{j,t}}{x_{j,0}}} = \frac{1}{L_{t/0}}$$

car

$$L_{t/0} = \sum_j c_{j,0} \frac{x_{j,t}}{x_{j,0}}$$

On a

$$P_{0/t} \neq \frac{1}{P_{t/0}}$$

L'indice de Paasche n'est pas réversible.

3.3.1.3 Indice de Fisher

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

On a

$$L_{0/t} = \frac{1}{P_{t/0}} \quad \text{et} \quad P_{0/t} = \frac{1}{L_{t/0}}$$

D'où

$$F_{0/t} = \sqrt{L_{0/t}\, P_{0/t}} = \frac{1}{\sqrt{L_{t/0}\, P_{t/0}}} = \frac{1}{F_{t/0}}$$

L'indice de Fisher est réversible.

3.3.2 Circularité

Aucun des trois indices ne possède la propriété de circularité.

$$\frac{I_{t/0}}{I_{t*/0}} = I_{t/t*}$$

Vérifions-le par exemple dans le cas d'un indice de Laspeyres des prix :

$$L_{t/0}(p) = \frac{\sum_j p_t q_0}{\sum_j p_0 q_0} \;;\; L_{t*/0}(p) = \frac{\sum_j p_{t*} q_0}{\sum_j p_0 q_0}$$

$$\frac{L_{t/0}(p)}{L_{t*/0}(p)} = \frac{\sum_j p_t q_0}{\sum_j p_{t*} q_0} \quad ; \quad L_{t/t*}(p) = \frac{\sum_j p_t q_{t*}}{\sum_j p_{t*} q_{t*}}$$

On a alors

$$\frac{L_{t/0}(p)}{L_{t*/0}(p)} \neq L_{t/t*}(p)$$

L'indice de Laspeyres n'est pas circulaire.

3.4 Relations entre les différents indices

3.4.1 Relation 1

L'indice de valeur est égal au produit de l'indice de Laspeyres des prix par l'indice de Paasche de quantités, ou inversement, c'est à dire :

$$\begin{aligned} V_{t/0} &= L_{t/0}(p)\,.\,P_{t/0}(q) \\ &= L_{t/0}(q)\,.\,P_{t/0}(p) \end{aligned}$$

En effet :

$$\begin{aligned} L_{t/0}(p).P_{t/0}(q) &= \frac{\sum_j p_t\,.q_0}{\sum_j p_0\,.q_0} \times \frac{\sum_j p_t\,.q_t}{\sum_j p_t\,.q_0} \\ &= \frac{\sum_j p_t\,.q_t}{\sum_j p_0\,.q_0} = V_{t/0} \end{aligned}$$

et

$$\begin{aligned} L_{t/0}(q).P_{t/0}(p) &= \frac{\sum_j p_0\, q_t}{\sum_j p_0\, q_0} \times \frac{\sum_j p_t\, q_t}{\sum_j p_0\, q_t} \\ &= \frac{\sum_j p_t\, q_t}{\sum_j p_0\, q_0} = V_{t/0} \end{aligned}$$

Exercice 7

Nous considérons les données de l'exercice 6 et donnons une interprétation des formules établies ci-dessus.

Formule 1

$$V_{t/0} = L_{t/0}(p)\; P_{t/0}(q)$$

$$1{,}3907 = 1{,}0967 \times 1{,}2681$$

L'augmentation de 39,07% de valeur est due à une augmentation de 9,67% des prix et à une augmentation de 26,81% des quantités.

Formule 2

$$V_{t/0} = L_{t/0}(q)\; P_{t/0}(p)$$

$$1{,}3907 = 1{,}3197 \times 1{,}0538$$

L'augmentation de 39,07% de valeur est due à une augmentation de 5,38% des prix et à une augmentation de 31,97% des quantités.

3.4.2 Relation 2

L'indice de valeur est égal au produit de l'indice de Fisher des prix par l'indice de Fisher des quantités.

$$V_{t/0} = \sqrt{V_{t/0}\ .\ V_{t/0}} \quad = \sqrt{L_{t/0}(p) \,.\, P_{t/0}(q) \times L_{t/0}(q) \,.\, P_{t/0}(p)}$$

$$= \sqrt{L_{t/0}(p) \,.\, P_{t/0}(p)} \times \sqrt{L_{t/0}(q) \,.\, P_{t/0}(q)} = F_{t/0}(p) \times F_{t/0}(q)$$

3.5 Les indices chaînes

Si $I_{t/t-1}$ est l'indice synthétique à la date t par rapport à la date t – 1, l'indice chaîne à la date t par rapport à la date 0 vaut :

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

L'indice obtenu possède la propriété de circularité. On pourra par conséquent raccorder les indices-chaînes s'ils sont définis de la même façon.

La formule permet de comparer la situation actuelle à celle de la période précédente.

Les difficultés se présentent lorsqu'on enchaîne les éléments successifs :

a. une erreur dans l'un des éléments de la chaîne se répercute dans les indices suivants,

b. le calcul est très long,

c. l'indice perd rapidement toute signification économique. Le budget de référence est modifié et le manque de signification apparaît d'autant plus évident que le nombre d'anneaux augmente.

Exercice 8 : Calcul d'un indice chaîne

$I_{1996/1995} = 105 \quad ; \quad I_{1997/1996} = 102$

$I_{1998/1997} = 106 \quad ; \quad I_{1999/1998} = 101$

$I_{1999/1995}$ est calculé en prenant pour base 100 l'année 1995. On obtient :

$$I_{1999/1995} = \frac{105 \times 102 \times 106 \times 101}{100 \times 100 \times 100}$$

$$I_{1999/1995} = 114{,}6$$

Bibliographie

Calot G, Cours de statistique descriptive, Paris, Dunod, 1973.

Doucouré F.B, Statistique descriptive appliquée aux sciences de santé : Cours et exercices corrigés avec les logiciels Eviews, Stata et SPSS, Presses Universitaires de Dakar (PUD), 2021.

Doucouré F.B, Statistique et probabilités pour économistes et gestionnaires ; Cours et exercices corrigés, Editions ARIMA, 2020.

Fourastié J, Laslier J. F., Probabilités et statistiques, Dunod, 1997

Grais B, Méthodes statistiques, Paris, Dunod, 1974.

Masiéri W, Statistique et calcul des probabilités, Paris, Sirey, 1994.

Tassi P, Méthodes statistiques, Economica, 1992

COFEB
CENTRE OUEST AFRICAIN DE FORMATION
ET D'ETUDES BANCAIRES

**Centre Ouest Africain de Formation
et d'Etudes Bancaires (COFEB)**

Avenue Abdoulaye Fadiga
BP : 3108 Dakar - Sénégal
Téléphone : 00 221 33 839 05 00
Fax : 00 221 33 823 83 35
Contact : courrier.zdrpbceao.int