République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : **Statistique**

Par

KHECHAI Ines

Titre :

La statistique descriptive univariée

Membres du Comité d'Examen :

Pr	BENATIA FATAH	UMKB	Président
Dr.	SAYAH ABDALLAH	UMKB	Encadreur
Dr.	TOUBA SONIA	UMKB	Examinateur

Juin 2021

Dédicace

Je dédie ce humble travail à :

♡♡ Mes chers parents ♡♡

Tous les membres de ma famille :

♡ Ma chère soeur "Imane", et son Mari "Mohamed" ♡

♡ Ma petite soeur "Rim" ♡

Mes troix frères :

♡ "Sofiane", "Mohamed Islam", "Ibrahim" ♡

Mes chères amies :

♡ Oumaima, Sabrine ♡

Tous les professeurs qui m'ont enseigné dans mes années précedentes

Salmoni Ali, Belmassaoud Azzeddine.

Tous les enseignants et les collegues de promotion de la deuxième année,

Master 2020 - 2021.

REMERCIEMENTS

Je tiens tout d'abord à remercier "Allah" le tout puissant de m'avoir aidé et

Donné la santé pour arriver à ce stade.

Je remercie les deux personnes les plus importantes de ma vie :

Ma mère et Mon père

Je ne peux jamais leur dire des mots qui leur rapportent ce qu'ils méritent,

Sans leur soutien

Je ne serais jamais ici, merci beaucoup à eux et qu'Allah bénisse leurs jours et santés.

Mes vifs remerciements et gratitude à mon encadreur : **Dr.SAYAH ABDALLAH**,

Qui m'a guidé avec soin et patience tout en faisant ce travail,

Et lui souhaite le meilleur dans sa carrière professionnelle.

Je tiens à remercier : **Pr. BENATIA FATAH**, **Dr. TOUBA SONIA**,

De m'avoir honoré et accepté d'évaluer ce travail.

Je remercie tous ce qui ont contribué dans ce travail,

De prés ou de loin.

Table des matières

Table des figures

Liste des tableaux

Introduction

La statistique est un ensemble de méthodes mathématiques basées sur l'organisation et la présentation de données ce qui conduit à la construction de résumé numérique, de décrire et d'analyser des phénoménes succeptibles d'être dénombrés. On distingue généralement deux types : **la statistique inférentielle** et **la statistique descriptive**, ce dernier vise à étudier et décrire de façon synthétique et parlante des données observées pour mieux les analyser, et qui à leur tour, sont composées en deux parties : la statistique descriptive univariée et la statistique descriptive multivariée.

Nous nous intéréssons dans notre mémoire à la statistique descriptive univariée et l'étude de données associées d'une seule variable, que celle-ci soit d'une variable qualitative ou quantitative.

Notre mémoire est divisé en trois chapitres :

Dans le **chapitre 1**, on introduit des généralités et des notions de base sur la statistique et la représentation des variables sous forme de tableaux et graphiques (secteur angulaire, L'histogramme...).

Le **chapitre 2**, est consacré à une présentation détaillée sur les caractéristiques (Les caractéristiques de tendance centrale, de dispersion et de forme) suivi par des exemples explicatifs.

Dans **le chapitre 3,** nous allons présenter quelques exemples concernants les deux chapitres précédents en utilisant le logiciel de programmation **R.**

Chapitre 1

Introduction et concepts de base

Ce chapitre contient des généralités et des concepts fondamentaux sur la statistique descriptive.

1.1 Définitons fondamentales

Définition 1.1.1 *: La Statistique est une discipline qui a pour objet la collecte, le traitement et l'analyse de données numériques relatives à un ensemble d'individus ou d'éléments, c'est aussi un ensemble de méthodes scientifiques dont l'objectif est d'analyser et modéliser des informations numériques.*

1.1.1 La statistique descriptive

Définition 1.1.2 *: La statistique descriptive est un ensemble de méthodes utilisées pour décrire les caractéristiques étudiées d'un ensemble de données à l'aide de moyens appropriés, et elle vise à décrire, classer, et organiser, résumer un ensemble de données (qualitatives,quantitatives), puis l'afficher clairement sous forme de tableaux en fonction des valeurs calculées (moyenne, médiane, écart type...) ou des graphiques (histogramme, camembert graphique...).*

Elle se compose de deux domaines distincts :

La statistique descriptive univariée : Correspond à l'analyse d'un seul caractère, c'est l'étude de la population selon une seule variable.

La statistique descriptive multivariée : C'est l'étude de la population à plusieurs variables. Par exemple la statistique descriptive bivariée (est un cas particulier à deux variables).

1.1.2 Notions de bases statistiques

Population : On appelle population l'ensemble des unités statistiques homogènes ou ensemble des éléments auxquels se rapportent les données étudiées, elle est notée Ω. Par exemple : les étudiants d'une classe, ensemble des habitants d'une ville...

Echantillon : On appelle échantillon le sous-ensemble de la population sur lequel sont effectivement réalisées les observations. Par exemple : l'ensemble des étudiants d'une salle de classe d'une université, l'ensemble d'un millier d'habitants choisi parmi tous les habitants d'une ville.

Individu : (Unité statistique) élément de base constituant la population ou l'échantillon, elle est notée ω. Par exemple : l'étudiant d'une université, le livre d'une bibliothèque.

Caractère ou variable statistique : (C'est la propriété étudiée) Un caractère X étant une variable qui discerne les individus de cette population Ω, les valeurs possibles d'un caractère sont appelées ses modalités.

Modalités : Les modalités x_i sont les différentes possibilités que peut prendre le caractère X (ou les différentes situations de X), où chaque caractère a deux ou plusieurs façons de modalités, par exemple :

1). Les modalités du caractère **sexe** sont **masculin** et **féminin**.

2). Les modalités du caractère **nationalité** sont **Algérien**, **Marocain**, **Français**,...

1.1.3 Type de Caractère (ou variable statistique)

Il existe deux types de caractères : Les caractères **qualitatifs** et les caractères **quantitatifs**.

Caractère qualitatif :

Définition 1.1.3 *: Un caractère est dit qualitatif si toutes ses valeurs possibles ne sont pas numériques, il ne peut pas être mesuré, et on distingue deux types de variables qualitatives :*

Variable qualitative nominale : *Une variable qualitative est dite nominale lorsque ses modalités ne peuvent être classées et ordonnées de façon naturelle. Par exemple : le cas de la variable couleur des yeux, ou encore de la variable de sexe, la nationalité...*

Variable qualitative ordinale : *Une variable qualitative est dite ordinale lorsque ses modalités peuvent être classées dans un certain ordre naturel. Par exemple : le degré de sévérité d'une maladie (forte, moyenne, faible).*

Caractère quantitatif :

Définition 1.1.4 *: Un caractère est dit quantitatif si toute ses valeurs possibles ou l'ensemble des observations sont numériques, et ses différentes modalités sont mesurables ou repérables, et on distingue deux types de variables quantitatives :*

Variable quantitative discrète : *Une variable est dite quantitative discrète, si l'ensemble des observations possibles est fini ou dénombrable, ou de façon générale entre deux valeurs successives aucune autre valeur n'est possible. Par exemple : le nombre de pages d'un livre, la valeur pointure des souliers,...*

Variable quantitative continue : *Elle est dite continue si le nombre de modalités est infini et ne sont pas des valeurs précises. Par exemple : La taille, le poids, la durée de vie d'un produit, le nombre de bactéries..., d'autre façon, il peut prendre* *toutes les valeurs d'un intervalle inclu dans* $\mathbb{R}$ *(* $[b_i, b_{i+1}[$ *).*

1.1.4 Effectifs, fréquences, fréquences cumulées

Effectif partiel :

Définition 1.1.5 *: On appelle effectif d'une modalité* x_i*, le nombre d'individus observés correspondant à chaque caractère* X *de la population* Ω*, on le note* n_i *:*

$$n_i = card\{\omega \in \Omega \quad X(\omega) = x_i\} \qquad\qquad i = 1, 2, ..., k. \tag{1.1}$$

Effectif total :

Définition 1.1.6 *: L'effectif total est le nombre total d'individus constituant la population statistique étudiée, c'est aussi la somme de tous les effectifs et il est noté* n *:*

$$n = \sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k = card(\Omega). \tag{1.2}$$

Fréquence partielle (fréquence relative) :

Définition 1.1.7 *: La fréquence partielle (relative) associée à une modalité est la proportion d'individus de la population totale appartenant à la classe, elle est égale au rapport de l'effectif de cette modalité* n_i *par l'effectif total* n*, on la note* f_i *:*

$$f_i = \frac{n_i}{n}. \tag{1.3}$$

Remarque 1.1.1 *: 1) La somme des fréquences relatives est égale à* 1 *car la valeur de la fréquence est toujours entre* 0 *et* 1*".*

$$\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i = \frac{1}{n}.n = 1.$$

2) Si l'on multiplie par 100 *les fréquences* $(100 \times f_i)$*, nous obtenons des* ***fréquences en pourcentage****, notées* $f_i(\%)$.

Effectif et fréquence cumulé croissant :

Définition 1.1.8 *: L'effectif cumulé croissant ECC (fréquence cumulée croissante FCC) d'une modalité est la somme des valeurs des tout effectif correspondantes n_i (ou fréquence f_i) aux valeurs de la variable statistique inférieures ou égales à x_i, comme suit :*

$$ECC = N_i = \sum_{j=1}^{i} n_j, \quad FCC = F_i = \sum_{j=1}^{i} f_j. \tag{1.4}$$

Effectif et fréquence cumulé décroissant :

Définition 1.1.9 *: L'effectif cumulé décroissant ECD (fréquence cumulée décroissante FCD) d'une modalité est la somme des valeurs des tout effectif correspondant n_i (ou fréquence f_i) aux valeurs de la variable statistique supérieures ou égales à x_i, comme suit :*

$$ECD = N_i' = \sum_{j=i}^{k} n_j, \qquad FCD = F_i' = \sum_{j=i}^{k} f_j. \tag{1.5}$$

1.2 Représentations des données

1.2.1 Série statistique

Définition 1.2.1 *: On appelle série statistique la suite des valeurs ou l'ensemble des données associées à un certain caractère au sein de cette population.*

1.2.2 Tableau statistique

Définition 1.2.2 *: Un tableau statistique est une méthode permettant de visualiser, classifer et ranger par ordre croissant (ou décroissant) des données. Il peut être aussi bien utilisé pour représenter des données brutes de la série statistique.*

Il existe **deux types de tableaux** selon la nature de la variable, si la variable est qualitative ou quantitative discrète, elle est classifée comme des valeurs numériques, et comme des intervalles si la variable est quantitative continue.

Tableau statistique d'une variable qualitative et quantitative discrète :

Modalités du caractère X	Effectifs n_i	$ECC\ N_i$	$ECD\ N_i'$
Modalités x_1	n_1	N_1	$N_1' = n_k + ... + n_1$
Modalités x_2	n_2	$N_2 = n_1 + n_2$	$N_2' = n_k + ... + n_2$
...	...	...	...
Modalités x_i	n_i	$N_i = n_1 + ... + n_i$	$N_i' = n_k + ... + n_i$
...	...	...	...
Modalités x_k	n_k	$N_k = n$	$N_k' = n_k$
Total	n		

Fréquences f_i	$FCC\ F_i$	$FCD\ F_i'$
$f_1 = \frac{n_1}{n}$	F_1	$F_1' = f_k + ... + f_1$
$f_2 = \frac{n_2}{n}$	$F_2 = f_1 + f_2$	$F_2' = f_k + ... + f_2$
...	...	...
$f_i = \frac{n_i}{n}$	$F_i = f_1 + f_2 + ... + f_i$	$F_i' = f_k + ... + f_i$
...	...	...
$f_k = \frac{n_k}{n}$	$F_k = f_1 + ... + f_i + ... + f_k = 1$	$F_k' = f_k$
1		

TAB. 1.1 – Tableau statistique d'une variable qualitative et quantitative discrète.

Tableau statistique d'une variable quantitative continue :

Les classes	c_i	a_i	n_i	f_i	$ECC\ N_i$	$FCC\ F_i$	$ECD\ N_i'$	$FCD\ F_i'$
$[b_1; b_2[$	c_1	a_1	n_1	f_1	N_1	F_1	N_1'	F_1'
$[b_2; b_3[$	c_2	a_2	n_2	f_2	N_2	F_2	N_2'	F_2'
...	...	...	...	...	...	...	...	...
$[b_k, b_{k+1}[$	c_k	a_k	n_k	f_k	N_k	F_k	N_k'	F_k'
Total			n	1				

TAB. 1.2 – Tableau statistique d'une variable quantitative continue.

- b_{i+1} et b_i : Les bornes d' une classe.
- $a_i = b_{i+1} - b_i$: L'amplitude d'une classe.
- $c_i = \dfrac{b_{i+1} + b_i}{2}$: Le centre d'une classe.

1.3 Représentation graphique des caractères

On représente généralement le caractère (qualitatif ou quantitatif), par des graphiques qui utilisent des surfaces. Il existe plusieurs graphiques pour chaque type de variable.

1.3.1 Cas d'une variable qualitative

On distingue plusieurs façons de représentations graphiques en ce qui concerne le caractère qualitatif, dont le plus courant sont les suivants : Représentation en **cercle** et **rectangle**.

Représentation par rectangle (Le diagramme en bandes) :

Dans cette représentation, les diférentes modalités du caractère sont représentées par des rectangles de même largeur, dont la base est constante (ne possède aucune sinification numérique), sa hauteur est proportionnelle à l'effectif (ou à la fréquence).

Représentation par cercle (Le diagramme en secteur angulaire) :

Représentation circulaire (ou semi-circulaire) consiste à diviser un disque (ou un demi-disque) en tranches ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, (ou à la fréquence), de la modalité, l'angle au centre β_i est égale à :

$$\beta_i = \begin{cases} 360 \times \dfrac{n_i}{n} = 360 \times f_i \Longrightarrow \text{Représentation circulaire.} \\ 180 \times \dfrac{n_i}{n} = 180 \times f_i \Longrightarrow \text{Représentation semi-circulaire.} \end{cases} \tag{1.6}$$

Exemple 1.3.1 *: (Variable qualitative nominale). Le tableau suivant donne la répartition selon le groupe sanguin de 50 individus pris au hasard dans une population :*

Modalités	n_i	f_i	$f_i(\%)$	$ECC\ N_i$	$ECD\ N_i'$	$FCC\ F_i$	$FCD\ F_i'$	angles β_i
A	12	$\frac{12}{50} = 0.24$	24%	12	50	0.24	1	86.4°
AB	8	$\frac{8}{50} = 0.16$	16%	20	38	0.40	0.76	57.6°
O	19	$\frac{19}{50} = 0.38$	38%	39	30	0.78	0.6	136.8°
B	11	$\frac{11}{50} = 0.22$	22%	50	11	1	0.22	79.2°
Total	50	1	100%					360°

TAB. 1.3 – Tableau représente nombre des personnes par le groupe sanguin .

Les angles dans le tableau se calcule en utilisant la règle de (1.6).

1. La population dans cette étude est : Les 50 personnes, la variable statistique est le groupe sanguin, elle est qualitative nominale.
2. Nous avons deux représentations possibles "diagramme en bandes" et "secteur angulaire".

kh 2021/memoire finale ines kh 2021/Mémoire Master 2021/QTKTQZ00.bmp

FIG. 1.1 – Les diagramme d'une variable qualitative.

1.3.2 Cas d'une variable quantitative

Dans ce cas, il existe deux types de représentation graphique d'une distribution statistique à caractère quantitatif :

Le diagramme différentiel :

Le diagramme différentiel est une représentation des effectifs ou des fréquences (absolues ou relatives).

Cas d'une variable quantitative discrète :

Le diagramme en bâtons : Un diagramme en bâtons est un graphique qui représente la distribution des effectifs ou des fréquences d'une variable discrète, ce diagramme comporte donc un axe horizontal (l'abscisse) qui représente les valeurs de la variable, et un axe vertical (l'ordonnée), dont la hauteur est proportionnelle à l'effectif (ou à la fréquence) correspondant.

Cas d'une variable quantitative continue :

L'histogramme : La représentation graphique lorsque les données sont regroupées en classes de même amplitude selon les fréqunences absolues ou relatives est appelée l'histogramme, la hauteur de chaque rectangle soit proportionnelle à la densité de l'effectif $n_i^{'}$ (l'effectif corrigé) ou la densité de fréquence $f_i^{'}$ (la fréquence corrigée).

$$n_i^{'} = \frac{n_i}{a_i}, \qquad f_i^{'} = \frac{f_i}{a_i}. \tag{1.7}$$

Remarque 1.3.1 *: 1) La surface de chaque rectangle est égale à l'effectif ($n_i^{'} \times a_i = n_i$) ou à la fréquence ($f_i^{'} \times a_i = f_i$) de la classe.*

2) Si on travaille avec les effectifs, l'aire de l'histogramme est égale l'effectif total n, ou égale 1, si on travaille avec les fréquences.

Le diagramme cumulatif (Le diagramme intégral) :

Le diagramme cumulatif correspond à une représentation graphique des effectifs et fréquences cumulés (croissants ou décroissants), est une courbe en escalier, les paliers horizontaux ont pour ordonnées ECC (ECD) ou FCC (FCD), ils sont obtenus à partir de la fonction de répartition $F(x)$ {voir (1.8)}.

Remarque 1.3.2 *: La fonction de répartition :*

$$F \qquad \begin{aligned} &\mathbb{R} \rightarrow [0,1] \\ &X \mapsto F(x) = P(X \leq x) \end{aligned} \tag{1.8}$$

Exemple 1.3.2 *: (Cas d'une variable quantitative discrète). On indique dans le tableau suivant la distance entre le bureau et le domicile (en km) d'un groupe d'employés d'une entreprise, on note les résultats suivants :*

Distance en (km)	1	2	3	4	5	6	7	**Total**
n_i	11	10	15	20	17	19	8	100
f_i	0.11	0.1	0.15	0.2	0.17	0.19	0.08	1
$ECC\ N_i$	11	21	36	56	73	92	100	
$ECD\ N_i'$	100	89	79	64	44	27	8	
$FCC\ F_i$	0.11	0.21	0.36	0.56	0.73	0.92	1	
$FCD\ F_i'$	1	0.89	0.79	0.64	0.44	0.27	0.08	

TAB. 1.4 – Tableau du distance entre le bureau et le domicile d'un groupe d' employés parisiens.

- La représentations possibles est :

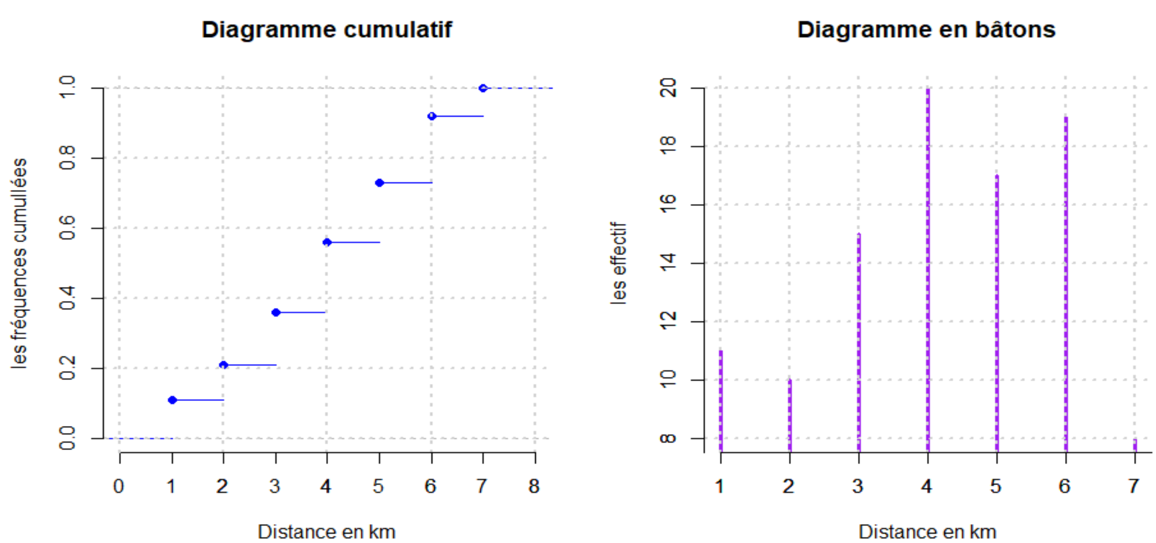

FIG. 1.2 – Les diagrammes d'une variable quantitative discrète.

Exemple 1.3.3 : *(Cas d'une variable quantitative continue d'amplitudes égales). Classement de 40 élèves selon le taux de réussite obtenu au baccalauréat :*

Les classes	$[10, 12[$	$[12, 14[$	$[14, 16[$	$[16, 18[$	**Total**
n_i	10	18	8	4	40
c_i	11	13	15	17	/
f_i	0.25	0.45	0.20	0.10	1
$ECC\ N_i$	10	28	36	40	/
$FCC\ F_i$	0.25	0.70	0.90	1	/
n_i'	5	9	4	2	/
f_i'	0.125	0.225	0.100	0.050	/

TAB. 1.5 – Tableau représente classement de 40 élève selon le taux de réussite obtenu au baccalauréat.

- La représentation possible est :

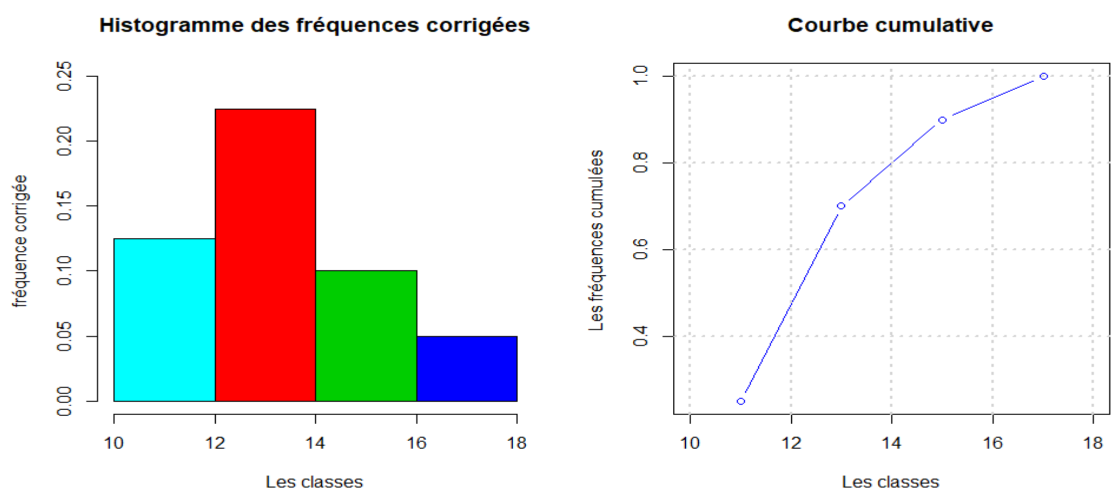

FIG. 1.3 – Les diagrammes d'une variable quantitative continue.

Chapitre 2

Statistique descriptive univariée

La Statistique descriptive univariée est une étude liée à une seule variable ou la description d'une série statistique selon un seul caractère, qui permer de présenter des données, à l'aide de tableaux statistiques ou de techniques graphiques et les résumés d'indicateurs numériques, le chapitre précédent traitait les données graphiquement, on veut maintenant étudier dans ce chapitre les paramètres caractéristiques.

2.1 Paramètres caractéristiques

Le but de l'étude statistique est décrire et résumer numériquement les données observées d'une variable des séries statistiques, ces descriptions sont obtenues à l'aide de paramètres caractéristiques, on peut proposer trois types de paramètres : Paramètres de position, paramètres de dispertion, et paramètres de forme.

2.1.1 Les caractéristiques de tendance centrale

Les caractéristiques de tendance centrale dit aussi paramètres de position donnent une information sur l'ordre de grandeur de la série statistique, ils est permettent de situer la valeur centrale autour du quel les données tournent.

Le mode :

Définition 2.1.1 *: Le mode ou valeur dominante d'une série statistique est la modalité ou l'observation correspondante à l'effectif (ou la fréquence) le plus élevé (la plus fréquente), on parle de* ***la classe modale*** *dans le cas d'une variable quantitaive continue, si les classes ont les même amplitudes (amplitudes égales), alors la classe modale correspond à l'effectif (ou la fréquence) le plus élevé, si les amplitudes de classes inégales, la classe modale est la densité d'effectif (ou la densité de fréquence) le plus élevé, est noté : M_o ou x_M et la valeur exacte du mode est :*

$$M_o = b_i + (b_{i+1} - b_i) \times \frac{(n'_{i+1} - n'_i)}{(n'_{i+1} - n'_i) + (n'_{i+1} - n'_{i+2})}. \tag{2.1}$$

- b_i et b_{i+1} : La borne inferieure et supérieure respectivement d'une classe.
- n'_i : La densité d'effectif précédente.
- n'_{i+1} : La densité d'effectif la plus élevée.
- n'_{i+2} : La densité d'effectif suivante.

Remarque 2.1.1 *: 1) Le mode peut ne pas exister, et s'il existe, il peut ne pas être unique.*

2) Le mode est calculé pour tous les types de variable (qualitative et quantitative).

3) Sur l'histogramme, la classe modale correspond au plus haut rectangle.

Exemple 2.1.1 *: On considère les séries de chiffres suivantes :*

1.

x_i	2	5	7	12
n_i	1	11	6	9

$\Leftrightarrow$ *Le mode de cette série est égal à 5, car il correspond au plus grand éffectif, qui vaut 11.*

2.

x_i	4	9	11	15
n_i	3	7	2	7

$\Leftrightarrow$ *Le mode de cette série est égal à 9 et 15, car il correspond respectivement aux plus grands éffectifs 7.*

3.

x_i	8	10	11	13
n_i	3	3	3	3

$\Leftrightarrow$ *Le mode de cette série n'existe pas, car toutes les valeurs ont le même éffectif* 3.

La médiane :

Définition 2.1.2 *: On appelle médiane (notée M_e ou $x_{\frac{1}{2}}$) la valeur de la variable statistique qui divise la population en deux sous-populations de même effectif, plus précisément, la médiane est la valeur centrale de la série d'observations, telle qu'il y ait autant d'individus supérieurs à M_e et d'individus inférieurs à M_e, la détermination de la médiane d'une série statistique nécessite d'abord de ranger par ordre croissant (ou décroissant) les valeurs observées, et on le calcule de la manière suivante :*

1- Variable quantitative discrète :

- Si $n = 2p$ observations, la médiane est la moyenne de p^{ieme} et de la $(p+1)^{ieme}$ observations, notée : $x_{(p)}$ et $x_{(p+1)}$.

$$M_e = \frac{1}{2} \times \left\{ x_{(p)+} x_{(p+1)} \right\} \tag{2.2}$$

- Si $n = 2p+1$ observations, la médiane est la valeur de la $(p+1)^{ieme}$ observation, notée : $x_{(p+1)}$.

$$M_e = x_{(p+1)} \tag{2.3}$$

2- Variable quantitative continue :

Dans le cas d'une variable continue, le calcul de la médiane passe, d'abord, par la détermination de la classe médiane ou classe qui contient la médiane, on précise la médiane par la formule suivante :

$$M_e = b_i + (b_{i+1} - b_i) \times \frac{\frac{n}{2} - N_i}{N_{i+1} - N_i} = b_i + a_i \times \frac{\frac{n}{2} - N_i}{N_{i+1} - N_i}. \tag{2.4}$$

- N_i : L'effectif cumulé croissant precédent.
- N_{i+1} : L'effectif cumulé croissant de la classe médiane.
- $\frac{n}{2}$: La moitié de l'effectif total.
- a_i : Amplitude de la classe médiane.

Remarque 2.1.2 *: 1) En remplaçant l'effectif cumulé par les fréquences, telle que $F_i < 0.5 < F_{i+1}$, la médiane est donnée par :*

$$M_e = b_i + (b_{i+1} - b_i) \times \left(\frac{0.5 \times F_{i+1}}{F_i - F_{i+1}}\right) = b_i + a_i \times \left(\frac{0.5 \times F_{i+1}}{f_i}\right) \tag{2.5}$$

2) La médiane existe pour tous les types de caractère, sauf le type d'un caractère qualitatif nominal.

Exemple 2.1.2 *: Soit la distribution statistique de données suivantes :*

- $\{0, 3, 5, 8, 11, 12, 17, 21, 25, 34\} \Longrightarrow n = 10 = 2 \times 5 \Longrightarrow$ La médiane est la moyenne de $x_{(5)}$ et $x_{(5+1)}$ (2.2):

$$M_e = \frac{1}{2} \times \left\{x_{(5)+}x_{(5+1)}\right\} = \frac{1}{2} \times \{11 + 12\} = 11.5.$$

- $\{6, 9, 10, 14, 18, 19, 23, 27, 30\} \Longrightarrow n = 9 = (2 \times 4) + 1 \Longrightarrow$ La médiane est la valeur de $x_{(4+1)}$ (2.3):

$$M_e = x_{(4+1)} = M_e = x_{(5)} = 18.$$

La moyenne arithmétique :

Définition 2.1.3 *: La moyenne arithmétique notée $\overline{X}$, est le quotient de la somme des valeurs observées sur leur nombre d'apparitions n (l'effectifs total) et elle est donnée par :*

*1). **Cas d'une variable statistique discrète :** Dans ce cas, on distingue deux types de moyenne :*

***La moyenne arithmétique simple** : On appelle moyenne arithmétique simple la somme de toutes les observations x_i (distinctes) divisée par le nombre de ces observations.*

$$\overline{X} = \frac{x_1 + x_2 + ... + x_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i. \tag{2.6}$$

***La moyenne arithmétique pondérée :** La moyenne arithmétique est dite pondérée, lorsque chaque modalité (x_i) est observée n_i fois.*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i \ , \qquad f_i = \frac{n_i}{n}. \tag{2.7}$$

*2). **Cas d'une variable statistique continue :** La distribution est présentée sous forme de classes, donc la formule de définition de la moyenne ne peut être appliquer directement, on prendra le centre de la classe, pour calculer la moyenne, donc :*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i \ , \qquad c_i \ \textit{: les centres de classes.} \tag{2.8}$$

Remarque 2.1.3 *: La moyenne ne se définit que pour un caractère statistique quantitatif (discrète ou continu).*

***Exemple 2.1.3** : (Cas d'une variable statistique discrète) Un élève a obtenu les notes suivantes :*

$x_i = \{4, 7, 12, 13.5, 12, 12, 10, 18, 13.5, 10, 11\} \Longrightarrow$ *La moyenne arithmétique vaut :*

- Dans le cas d'une **moyenne arithmétique simple** :

$$\begin{aligned}\overline{X} = \frac{1}{n}\sum_{i=1}^{k} x_i &= \frac{4+7+12+13.5+12+12+10+18+13.5+10+11}{11}, \\ &= \frac{123}{11} = 11.182.\end{aligned}$$

- Dans le cas d'une **moyenne arithmétique pondérée** :

x_i	4	7	10	11	12	13.5	18
n_i	1	1	2	1	3	2	1

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{(4\times 1)+(7\times 1)+(10\times 2)+(11\times 1)+(12\times 3)+(13.5\times 2)+(18\times 1)}{11},$$

$$= 11.182.$$

Propriétés fondamentales de la moyenne arithmétique :

Propriété 2.1.1 *: 1). La somme des ècarts à la moyenne arithmétique est nulle :*

$$\sum_{i=1}^{k} n_i(x_i - \overline{X}) = 0. \tag{2.9}$$

*2). **Transformation linéaire :** On définit Z, telle que : $z_i = ax_i + b$, a et b sont des constantes quelconques, alors la moyenne de Z obtenue à partir de la moyenne de X est :*

$$\overline{Z} = a\overline{X} + b.$$

Preuve. 1). [9]

$$\sum_{i=1}^{k} n_i(x_i - \overline{X}) = \sum_{i=1}^{k} n_i x_i - \sum_{i=1}^{k} n_i\overline{X} = \sum_{i=1}^{k} n_i x_i - \overline{X}\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} n_i x_i - n\overline{X}.$$

Et comme $\sum\limits_{i=1}^{k} n_i x_i = n\overline{X}$.

$$\sum_{i=1}^{k} n_i(x_i - \overline{X}) = n\overline{X} - n\overline{X} = 0.$$

2). [6], [10]

$$\begin{aligned}
\overline{Z} &= \frac{1}{n}\sum_{i=1}^{k} n_i z_i = \frac{1}{n}\sum_{i=1}^{k} n_i(ax_i + b),\\
&= \frac{1}{n}\sum_{i=1}^{k} n_i a x_i + \frac{1}{n}\sum_{i=1}^{k} n_i b, = \frac{a}{n}\sum_{i=1}^{k} n_i x_i + \frac{b}{n}\sum_{i=1}^{k} n_i,\\
\overline{Z} &= \frac{a}{n}\sum_{i=1}^{k} n_i x_i + b = a \times (\frac{1}{n}\sum_{i=1}^{k} n_i x_i) + b.
\end{aligned}$$

Alors :

$$\overline{Z} = a\overline{X} + b.$$

■

Propriétés générales de la moyenne arithmétique :

1- Si l'on multiplie par un scalaire λ chaque valeur de la série **alors** la moyenne arithmétique aussi est multipliée par λ [3] :

$$\frac{1}{n}\sum_{i=1}^{k} n_i(\lambda \times x_i) = \lambda \times \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \lambda \times \overline{X}.$$

2- Si on ajoute ou on retranche un nombre réel b à chaque valeur de la série **alors** on a [3] :

$$\frac{1}{n}\sum_{i=1}^{k} n_i(x_i \pm b) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \pm \frac{1}{n}\sum_{i=1}^{k} n_i b = \overline{X} \pm b.$$

Exemple 2.1.4 *: En prenant un exemple sur la répartition des 60 individus selon leur âge par le tableau suivant :*

Âge (ans)	n_i	c_i	a_i	f_i	$f_i(\%)$	N_i	$n_i^{'}$
$[18, 20[$	12	19	2	0.2	20	12	6
$[20, 23[$	21	21.5	3	0.35	35	33	7
$[23, 26[$	18	24.5	3	0.3	30	51	6
$[26, 28[$	9	27	2	0.15	15	60	4.5
Total	60	/	/	1	100	/	/

TAB. 2.1 – Tableau représente la répartition des 60 individus selon leur âge.

On va calculer les paramètres de position (le mode, la médiane, la moyenne) :

1. Le mode *:*

La classe modale correspond à l'effectif corrigé le plus grand $n_i^{'} = 7$, elle est égale à $[20, 23[$, on utilise la formule (2.1) :

$$\begin{aligned} M_o &= b_i + (b_{i+1} - b_i) \times \frac{(n_{i+1}^{'} - n_i^{'})}{(n_{i+1}^{'} - n_i^{'}) + (n_{i+1}^{'} - n_{i+2}^{'})}, \\ &= 20 + (23 - 20) \times \frac{(7-6)}{(7-6)+(7-6)}, \\ &= 20 + (3 \times \frac{1}{2}) = 21.5. \end{aligned}$$

La valeur modale est égale à 21.5, c'est celle qui correspond au plus grand âge de la distribution.

2. La médiane *:*

On a $n = 60$ (pair) et $\frac{n}{2} = 30$, alors la classe médiane est la classe qui contient la $30^{iéme}$ valeur et la $31^{iéme}$ valeur, d'aprés la colonne N_i du tableau, on cherche l'intervalle qui a un $N_i \geq 31$, on trouve que la classe médiane $[20, 23[$, on utilise (2.4) :

$$\begin{aligned} M_e &= b_i + (b_{i+1} - b_i) \times \frac{\frac{n}{2} - N_i}{N_{i+1} - N_i}, \\ &= 20 + (23 - 20) \times \frac{30 - 12}{33 - 12}, \\ &= 20 + (3 \times \frac{6}{7}) \backsimeq 22.571. \end{aligned}$$

3. La moyenne :

$$\begin{aligned} \overline{X} &= \frac{1}{n} \sum_{i=1}^{k} n_i c_i = \frac{1}{60} \left[(12 \times 19) + (21 \times 21.5) + (18 \times 24.5) + (9 \times 27) \right], \\ &= \frac{1363.5}{60} = 22.725. \end{aligned}$$

2.1.2 Les caractéristiques de dispersion

Les paramètres de dispersion étudient la dispersion des données autour de la valeur centrale, et quantifient les fluctuations des valeurs observées et leurs étalements.

L'étendue (range) :

Définition 2.1.4 *: L'étendue ou intervalle de variation d'une série statistique est l'écart entre la plus grande valeur et la plus petite valeur du caractère, elle est notée E :*

$$E = \begin{cases} x_{\max} - x_{\min} \Longrightarrow \text{ dans le cas discrète.} \\ c_k - c_1 \Longrightarrow \text{ dans le cas continue.} \end{cases} \tag{2.10}$$

Les quantiles :

Définition 2.1.5 *: Soit α un nombre compris entre 0 et 1, alors les quantiles d'ordre α sont les valeurs qui partagent la population totale en α parties égales, en d'autre terme sont la solution de l'équation $F(x) = \alpha$ {voir (1.8)}.*

Il existe trois types de quantiles, lorsqu'on divise également la population par quatre, par dix ou par cent on obtient respectivement des **quartiles**, des **déciles** et des **centiles** :

Les quartiles : On définit trois quartiles (Q_1, Q_2, Q_3) qui sont les quartiles d'ordre $\frac{1}{4}$ (0.25), d'ordre $\frac{2}{4}$ (0.50), et d'ordre $\frac{3}{4}$ (0.75) respectivement, et on a :

$$F(Q_1) = 0.25, F(Q_2) = 0.5, F(Q_3) = 0.75.$$

Les déciles : $(D_1, ..., D_9)$ sont les neuf valeurs qui divisent le nombre total des observations en 10 effectifs égaux, et on a :

$$F(D_1) = 0.1, F(D_2) = 0.2, ..., F(D_9) = 0.9.$$

Les centiles : $(C_1, ..., C_{99})$ sont les 99 valeurs qui divisent le nombre total des observations en 100 effectifs égaux, et on a :

$$F(C_1) = 0.01, F(C_2) = 0.02, ..., F(C_{99}) = 0.99.$$

Remarque 2.1.4 *: Le quartile se détermine de la même manière que la médiane.*

Les intervalles interquantiles :

Définition 2.1.6 *: L'intervalle interquantile est la distance entre le premier et les dernier quantile calculé, il est définie par la quantité :*

- **Intervalle interquartile :** $IQ = (Q_3 - Q_1)$ qui contient 50% des observations.
- **Intervalle interdécile :** $ID = (D_9 - D_1)$ qui contient 80% des observations.
- **Intervalle intercentile :** $IC = (C_{99} - C_1)$ qui contient 98% des observations.

Boîte à moustache (boxplot)

Définition 2.1.7 *: La boîte à moustache est un outil graphique sous la forme d'un rectengle, où chaque arête représente respectivment Q_1, Q_3, et Q_2 ou la médiane M_e qui divise le rectangle en deux parties, en plus de spécifier la valeur minimale et maximale de la série statistique.*

Exemple 2.1.5 *: Soit la distribution statistique de données suivantes :*

$x_i = \{4, 6, 7, 11, 12, 15, 19, 20, 21, 30, 33\}$, *alors les quartiles valent :*

- $Q_1 = x_{(3)} = 7$, $\quad Q_2 = x_{(6)} = 15.$ *et* $Q_3 = x_{(9)} = 21.$

- *Intervalle interquartile est* : $IQ = (Q_3 - Q_1) = 21 - 7 = 7.$

- ***Diagramme en boîte à moustaches :***

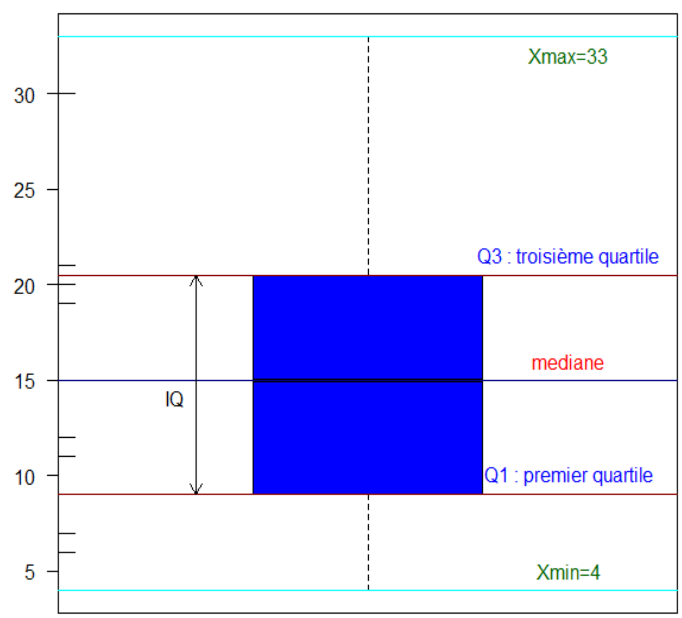

FIG. 2.1 – Boîte à moustache.

L'écart absolu moyen :

Définition 2.1.8 *: L'écart absolu moyen, noté e_{moy}, est la moyenne de la différence entre les valeurs observées et leur moyenne arithmétique, en valeur absolue, il est défini comme suit :*

$$e_{moy}(X) = \frac{1}{n}\sum_{i=1}^{k} n_i \left| x_i - \overline{X} \right| = \sum_{i=1}^{k} f_i \left| x_i - \overline{X} \right|. \tag{2.11}$$

Propriété 2.1.2 *1). On a $e_{moy}(X) \geq 0$. De plus on a l'équivalence suivante :*

$$e_{moy}(X) = 0 \Longleftrightarrow x_1 = x_2 = ... = x_k.$$

2). Soit a et b deux constantes réelles, si Z est une transformation de X donnée par $Z = aX + b$ alors :

$$e_{moy}(Z) = |a| \times e_{moy}(X).$$

Preuve. [12]

1). Comme $e_{moy}(X)$ est une somme de valeurs absolues divisée par $n > 0$, et l'écart absolu moyen est un rapport de deux nombres non-négatifs, donc il est non-négatif.

$$e_{moy}(X) = 0 \Longleftrightarrow \frac{1}{n}\sum_{i=1}^{k} n_i \left| x_i - \overline{X} \right| = 0 \Longleftrightarrow \left| x_i - \overline{X} \right| = 0,\ \forall i = \overrightarrow{1, ..., k},$$

où la dernière équivalence exprime le fait que la somme à termes positifs ou nuls est nulle si et seulement si tous ses termes sont nuls. La dernière condition est équivalente à $x_1 = x_2 = ... = x_k$.

2). On a $\overline{Z} = a\overline{X} + b$ et $z_i = ax_i + b$,$\forall\ i = \overrightarrow{1, ..., k}$, par conséquent, $z_i - \overline{Z} = a(x_i - \overline{X})$ et :

$$e_{moy}(Z) = \frac{1}{n}\sum_{i=1}^{k} n_i \left|a(x_i - \overline{X})\right| = \frac{1}{n}\sum_{i=1}^{k} |a| \times n_i \left|(x_i - \overline{X})\right| = |a| \times \frac{1}{n}\sum_{i=1}^{k} n_i \left|(x_i - \overline{X})\right|,$$
$$= |a| \times e_{moy}(X).$$

■

La variance :

Définition 2.1.9 *: La variance est la moyenne des carrés de différence entre les valeurs observées et leur moyenne arithmétique, notée $Var(X)$ ou σ_X^2, donnée par la quantité :*

$$Var(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{k} f_i(x_i - \overline{X})^2. \quad (2.12)$$

Théor■me 2.1.1 *: (**Formule de KÖnig-Huygens**)*

Cette formule dit que la variance est la différence entre la moyenne des carrés de la variable X et le carrés de la moyenne :

$$Var(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{X}^2 = \sum_{i=1}^{k} f_i x_i^2 - \overline{X}^2. \quad (2.13)$$

Preuve. [4]

$$Var(X) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{X})^2 = \frac{\sum\limits_{i=1}^{k} n_i(x_i - \overline{X})^2}{\sum\limits_{i=1}^{k} n_i} = \frac{\sum\limits_{i=1}^{k} n_i(x_i^2 + \overline{X}^2 - 2x_i\overline{X})}{\sum\limits_{i=1}^{k} n_i},$$

$$= \frac{\sum\limits_{i=1}^{k} n_i x_i^2}{\sum\limits_{i=1}^{k} n_i} + \frac{\sum\limits_{i=1}^{k} n_i\overline{X}^2}{\sum\limits_{i=1}^{k} n_i} - 2 \times \frac{\sum\limits_{i=1}^{k} n x_i\overline{X}}{\sum\limits_{i=1}^{k} n_i},$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 + \frac{\sum\limits_{i=1}^{k} n_i\overline{X}^2}{\sum\limits_{i=1}^{k} n_i} - \frac{2\overline{X}}{n}\sum_{i=1}^{k} n_i x_i.$$

Ce qui implique que :

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

■

Propriété 2.1.3 *1). Si Z est une transformation de X donnée par :*

$$Z = aX + b \Longrightarrow Var(Z) = a^2 \times Var(X).$$

2). $Var(X) \geq 0$. On a l'égalité $Var(X) = 0 \Longleftrightarrow x_1 = x_2 = ... = x_k$.

Preuve. [12] 1). On a $\overline{Z} = a\overline{X} + b$ et $z_i = ax_i + b$, $\forall\ i = \overrightarrow{1, ..., k}$, par conséquent, $z_i - \overline{Z} = a(x_i - \overline{X})$ et :

$$Var(Z) = \frac{1}{n}\sum_{i=1}^{k} n_i\left[a(x_i - \overline{X})\right]^2 = \frac{1}{n}\sum_{i=1}^{k} a^2 n_i(x_i - \overline{X})^2,$$

$$= \frac{a^2}{n}\sum_{i=1}^{k} n_i(x_i - \overline{X})^2 = a^2 \times \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{X})^2 = a^2 \times Var(X).$$

2). Comme $Var(X)$ est une somme de carré divisée par $n > 0$, et la variance est un rapport de deux nombres non-négatifs, elle est donc non-négative.

$$Var(X) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{X})^2 = 0 \Longleftrightarrow \sum_{i=1}^{k} n_i(x_i - \overline{X})^2 = 0 \Longleftrightarrow (x_i - \overline{X})^2 = 0, \quad \forall\ i = \overrightarrow{1, ..., k}.$$

où la dernière équivalence exprime le fait que la somme, à termes positifs ou nuls, est nulle si et seulement si tous ses termes sont nuls. La dernière condition est équivalente à $x_1 = x_2 = ... = x_k$.[12] ■

L'écart-type :

Définition 2.1.10 *: L'écart-type notée σ_X, est égal la racine carrée positive de la variance, définie par La quantité :*

$$\sigma_X = \sqrt{Var(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{X})^2}. \tag{2.14}$$

Coefficient de variation :

Définition 2.1.11 *: On appelle coefficient de variation et on le note $CV(X)$ le rapport de l'écart type de X par rapport à sa moyenne arithmétique, et il est défini comme suit :*

$$CV(X) = \frac{\sigma_X}{|\overline{X}|}, \qquad CV(X) > 0. \tag{2.15}$$

Remarque 2.1.5 *: 1) Le Coefficient de variation est un nombre sans dimension, il sert à rendre les comparaisons entre les distributions statistiques différentes.*

2) Le Coefficient de variation, au signe près, est l'écart-type de la variable statistique $(\frac{X}{\overline{X}})$.

Moments :

Définition 2.1.12 *: On appelle moment d'ordre p (entier positif), par rapport à une valeur quelconque a (origine du moment), notée m_a^p, la quantité donnée par la formule suivante :*

$$m_p = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - a)^p = \sum_{i=1}^{k} f_i(x_i - a)^p. \tag{2.16}$$

On peut définir deux types de moments en fonction de a :

1- Les moments non centrés : Si la valeur de a est nulle ($a = 0$), on peut écrire m_a^p de la façon suivante :

$$m_p = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^p = \sum_{i=1}^{k} f_i x_i^p. \tag{2.17}$$

- Pour $p = 0 \Longrightarrow m_0 = 1$.
- Pour $p = 1 \Longrightarrow m_1 = \sum_{i=1}^{k} f_i x_i = \overline{X}$.
- Pour $p = 2 \Longrightarrow m_2 = \sum_{i=1}^{k} f_i x_i^2 = \overline{X}^2 = Q^2$ c'est la moyenne quadratique tel que : $Q = (\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2)^{\frac{1}{2}}$.

2- Les moments centrés : Si la valeur de a est égale à la moyenne arithmétique ($a = \overline{X}$), on peut écrire m_a^p de la façon suivante :

$$\mu_{\overline{X}}^p = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{X})^p = \sum_{i=1}^{k} f_i(x_i - \overline{X})^p. \tag{2.18}$$

- Pour $p = 0 \Longrightarrow \mu_{\overline{X}}^0 = 1$.
- Pour $p = 1 \Longrightarrow \mu_{\overline{X}}^1 = 0$.
- Pour $p = 2 \Longrightarrow \mu_{\overline{X}}^2 = Var(X)$.

Relations entre les moment centrés et les moment non centrés :

$$\begin{cases} \mu_X^2 = m_2 - (m_1)^2. \\ \mu_X^3 = m_3 - 3m_1m_2 + 2(m_1)^3. \\ \mu_X^4 = m_4 - 4m_1m_3 + 6(m_1)^2m_2 - 3(m_1)^4. \end{cases}$$

Exemple 2.1.6 *: D'aprés les données de l'exemple (2.1.4) on détermine les paramètres de dispersion suivants :*

1. L'étendue :

$$E = c_k - c_1 = 27 - 19 = 8.$$

2. Les quantiles :

On a $n = 60 \Longrightarrow \frac{n}{4} = 15$, à partir de la colonne des N_i, on cherche la valeur $N_i = 15$, tel que la classe qui correspond à cette valeur est $[20, 23[$, et en utilisant la relation suivante on a la valeur exacte de Q_1

$$\begin{aligned} Q_1 &= b_i + (b_{i+1} - b_i) \times \frac{\frac{n}{4} - N_i}{N_{i+1} - N_i}, \\ &= 20 + (23 - 20) \times \frac{15 - 12}{33 - 12}, \\ &= 20 + \frac{3}{7} \backsimeq 20.428. \end{aligned}$$

- $Q_2 = M_e = 22.571$.

La valeur exacte de Q_3 est $3 \times \frac{n}{4} = 45$, et la classe correspondante à $N_i = 45$ est $[23, 26[$, donc :

$$\begin{aligned} Q_3 &= b_i + (b_{i+1} - b_i) \times \frac{3 \times \frac{n}{4} - N_i}{N_{i+1} - N_i}, \\ &= 23 + (26 - 23) \times \frac{45 - 33}{51 - 33}, \\ &= 23 + (3 \times \frac{2}{3}) = 25. \end{aligned}$$

- Même principe pour les déciles et les centiles.

3. Intervalle interquantile :

$$IQ = (Q_3 - Q_1) = 25 - 20.428 = 4.572.$$

4. L'écart absolu moyen :

$$e_{moy}(X) = \frac{1}{n}\sum_{i=1}^{k} n_i \left|c_i - \overline{X}\right| = \frac{1}{60} \times [12 \times |19 - 22.725| + 21 \times |21.5 - 22.725|\,,$$

$$+\, 18 \times |24.5 - 22.725| + 9 \times |27 - 22.725|],$$

$$= \frac{1}{60} \times 140.85 = 2.3475.$$

5. La variance :

$$Var(X) = \frac{1}{n}\sum_{i=1}^{k} n_i (c_i - \overline{X})^2 = \frac{1}{60}[12 \times (19 - 22.725)^2 + 21 \times (21.5 - 22.725)^2,$$

$$+18 \times (24.5 - 22.725)^2 + 9 \times (27 - 22.725)^2],$$

$$= 7.105.$$

6. L'écart-type :

$$\sigma_X = \sqrt{Var(X)} = \sqrt{7.105} = 2.665.$$

7. Le coeffcient de variation :

$$CV(X) = \frac{\sigma_X}{\left|\overline{X}\right|} = \frac{2.665}{22.725} \backsimeq 0.117 = 11.7\%\,.$$

- On observe que l'écart-type, la variance et le coefficient de variation sont petits, qui signifie que la dispersion de la distribution est faible.

2.1.3 Les caractéristiques de forme

Les caractéristiques de forme sont des caractéristiques définies pour une variable statistique quantitative, elles permettent de préciser et décrire la forme de la distribution statistique ou de la courbe des fréquences par le coefficient d'asymétrie et d'aplatissement.

Coefficient d'asymétrie (skewness) :

Définition 2.1.13 *: Le Coefficient d'asymétrie a pour objet est savoir la façon régulière ou non dont les observations se répartissent de part et d'autre (obliquité à gauche ou à droite par rapport à une valeur centrale), et compare l'écart de la distribution par rapport à l'asymétrie.*

L'asymétrie d'une distribution par rapport aux trois valeurs de tendance centrale :

$$\begin{cases} \text{Lorsque } \overline{X} < M_e < M_o \Longrightarrow \text{on dit que la distribution est étalée à gauche.} \\ \text{Lorsque } \overline{X} = M_e = M_o \Longrightarrow \text{on dit que la distribution est symétrique.} \\ \text{Lorsque } \overline{X} > M_e > M_o \Longrightarrow \text{on dit que la distribution est étalée à droite.} \end{cases}$$

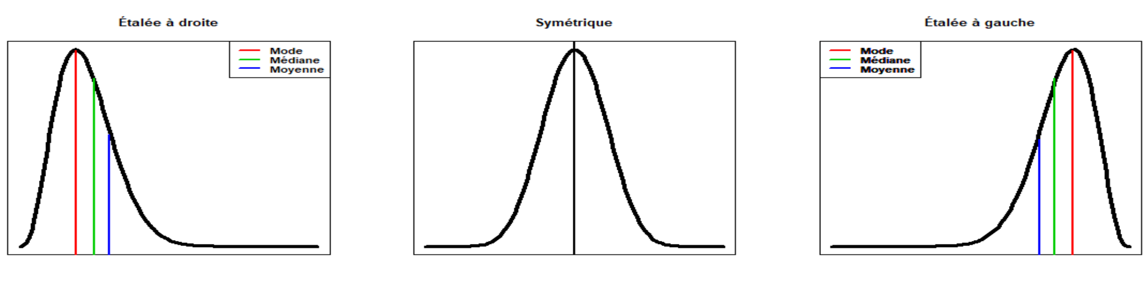

FIG. 2.2 – L'asymétrie d'une distribution.

Il existe plusieurs coefficients d'asymétrie, on peut proposer les principaux qui sont les suivants :

Le coefficient de Yule :

Définition 2.1.14 *: Le coefficient de Yule est calculé à partir de la position relative des quartiles* (Q_1, Q_2, Q_3), *ce coefficient est défini comme suit :*

$$Y = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)} = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1}. \tag{2.19}$$

$$\begin{cases} \text{Pour } Y < 0 \implies \text{la distribution est étalée à gauche,} \\ \text{pour } Y = 0 \implies \text{la distribution est symétrique,} \\ \text{pour } Y > 0 \implies \text{la distribution est a étalée à droite.} \end{cases}$$

Le coefficient de Pearson :

Définition 2.1.15 *: Le coefficient de Pearson est basé sur l'analyse de la position de la moyenne arithmétique et le mode. Il s'écrit de la forme suivante :*

$$P = \frac{\overline{X} - M_o}{\sigma_X}. \tag{2.20}$$

Aussi, il est défini à partir des moments centrés d'ordre 2 *et* 3 *comme suit :*

$$P = \frac{(\mu^3_{\overline{X}})^2}{(\mu^2_{\overline{X}})^3}. \tag{2.21}$$

$$\begin{cases} \text{Pour } P < 0 \implies \text{la distribution est étalée à gauche (oblique à droite),} \\ \text{pour } P = 0 \implies \text{la distribution est symétrique,} \\ \text{pour } P > 0 \implies \text{la distribution est étalée à droite (oblique à gauche).} \end{cases}$$

Le coefficient de Fisher :

Définition 2.1.16 *: Le coefficient de Fisher est basé sur le moment centré d'ordre trois, et il est la racine carrée du coefficient de Pearson, notée F, il s'écrit :*

$$F = \sqrt{\frac{(\mu_{\overline{X}}^3)^2}{(\mu_{\overline{X}}^2)^3}} = \frac{(\mu_{\overline{X}}^3)}{(\mu_{\overline{X}}^2)^{\frac{3}{2}}} = \frac{(\mu_{\overline{X}}^3)}{\sigma_X^3}. \tag{2.22}$$

$$\begin{cases} \text{Pour } F < 0 \implies \text{la distribution est étalée à gauche (oblique à droite),} \\ \text{pour } F = 0 \implies \text{la distribution est symétrique,} \\ \text{pour } F > 0 \implies \text{la distribution est étalée à droite (oblique à gauche).} \end{cases}$$

Coefficient d'aplatissement (Kurtosis) :

Définition 2.1.17 *: Ces coefficients ont pour but de comparer et mesurer l'aplatissement de la distribution présente à la l'aplatissement de la distribution normale.*

Les formes d'aplatissement : Il existe 3 formes de courbes d'aplatissement définies comme suit :

· **Aplatissement leptokurtique** (ou hyponormale) **:** Si la courbe est plus aplatie que la courbe normale.

· **Aplatissement normale** (mesokurtique) **:** Si la courbe présente un aplatissement normale.

· **Aplatissement platykurtique** (ou hypernormale) **:** Si la courbe est plus pointue que la courbe normale.

Voici une comparaison des trois situations :

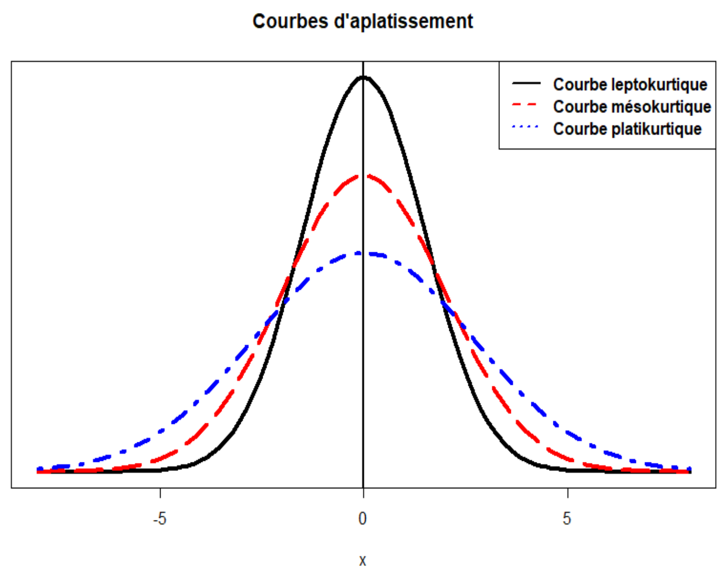

FIG. 2.3 – Aplatissements comparés.

On peut proposer deux Coefficients d'aplatissement :

Coefficient d'aplatissement de Pearson :

$$\beta_2 = \frac{\mu_X^4}{(\mu_X^2)^2} = \frac{\mu_X^4}{\sigma_X^4}. \tag{2.23}$$

$$\begin{cases} \text{Pour } \beta_2 < 3 & \Longrightarrow \text{la distribution est aplatie,} \\ \text{pour } \beta_2 = 3 & \Longrightarrow \text{la distribution est normale,} \\ \text{pour } \beta_2 > 3 & \Longrightarrow \text{la distribution est pointue.} \end{cases}$$

Coefficient d'aplatissement de Fisher : Dans le cas d'une distribution normale $\beta_2 = 3$, alors Fisher définit l'aplatissement par $\beta_2 - 3$.

$$F_2 = \beta_2 - 3 = \frac{\mu_X^4}{(\mu_X^2)^2} - 3. \tag{2.24}$$

$$\begin{cases} \text{Pour } F_2 < 0 \implies \text{la distribution est aplatie,} \\ \text{pour } F_2 = 0 \implies \text{la distribution est normale,} \\ \text{pour } F_2 > 0 \implies \text{la distribution est pointue.} \end{cases}$$

Exemple 2.1.7 *: D'aprés les données de l'exemple (2.1.4) on détermine les paramètres de forme suivants :*

1. Coefficient d'asymétrie :

- *Coefficient d'asymétrie de Pearson :*

$$P = \frac{\overline{X} - M_o}{\sigma_X} = \frac{22.725 - 21.5}{2.665} \backsimeq 0.459.$$

- *Coefficient d'asymétrie de Yule* :

$$Y = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1} = \frac{20.428 + 25 - 2(22.571)}{4.572} \backsimeq 0.0625.$$

- *Coefficient d'asymétrie de Fisher :*

$$\begin{aligned} \mu_{\overline{X}}^3 &= \sum_{i=1}^{4} f_i(c_i - \overline{X})^3 = [0.2 \times (19 - 22.725)^3 + 0.35 \times (21.5 - 22.725)^3 + 0.30(24.5 - 22.725)^3 \\ &\quad + 0.15 \times (27 - 22.725)^3] \\ &= 2.416\,2. \\ F &= \frac{(\mu_{\overline{X}}^3)}{\sigma_X^3} = \frac{2.416\,2}{(2.665)^3} = 0.12766. \end{aligned}$$

Comme $P > 0$ et $Y > 0$ et $F > 0$ et aussi on a " $\overline{X} > M_e > M_o$"$\Longrightarrow$ La distribution est étalée à droite (oblique à gauche).

2. Coefficient d'aplatissement :

- *Coefficient d'aplatissement de Pearson :*

$$\mu_{\overline{X}}^{4} = \sum_{i=1}^{4} f_i(c_i - \overline{X})^4 = [0.2 \times (19 - 22.725)^4 + 0.35 \times (21.5 - 22.725)^4 + 0.30(24.5 - 22.725)^4,$$

$$+ 0.15 \times (27 - 22.725)^4],$$

$$= 92.372.$$

$$\beta_2 = \frac{\mu_{\overline{X}}^{4}}{\sigma_X^4} = \frac{92.372}{(2.665)^4} = 1.8313.$$

- *Coefficient d'aplatissement de Fisher :*

$$F_2 = \beta_2 - 3 = 1.8313 - 3 = -1.1687.$$

Comme $\beta_2 < 3$ et $F_2 < 0 \Longrightarrow$ la distribution est aplatie.

Chapitre 3

Application sous R

Ce chapitre est une partie pratique dans le langage de prgrammation **R,** où nous résoudrons quelques exemples qui incluent ce qu'on a parlé dans les deux chapitres précédents, et nous traitons les paramètres de position et de dispersion et les représentations graphiques à l'aide du logiciel d'analyse statistique **R.**

3.1 Exemple sur la variable qualitative

Exemple 3.1.1 *: (Variable qualitative ordinale). Notre exemple est basé sur 80 enfants qui ont été interrogés sur l'addiction au chocolat, les résultats sont présentés dans le tableau suivant :*

Modalités	Pas du tout	Un peu	Beaucoup	Passionnément	**Totale**
Nombre des enfants	8	20	30	22	**80**

TAB. 3.1 – Tableau représente le nombre d'enfants ont été interrogées sur addiction au chocolat.

1. Calculer les f_i, ECC, FCC, le mode.

2. Tracer le diagramme en bandes et secteur angulaire.

Solution 3.1.1 *:*

Lecture des données :

x=c(rep("Pas du tout",8),rep("Un peu",20),rep("Beaucoup",30),rep("Passionnément",22))

y=c(table(x))

1. Le logiciel R nous donne les résultats suivants :

	n_i	ECC	f_i	FCC
Code R	*y*	*cumsum(y)*	*y/sum(y)*	*cumsum(y/sum(y))*
Beaucoup	30	30	0.375	0.375
Pas du tout	8	38	0.100	0.475
Passionnément	22	60	0.275	0.750
Un peu	20	80	0.250	1.000

· *Le mode = Beaucoup.*

2. La représentation graphique : Instruction R.

- ***Pour tracer le diagramme en bandes :*** *On utilise la fonction barplot(y).*

- ***Pour tracer le diagramme circulaire :*** *On utilise la fonction pie(y).*

Commentaire : *Pour le diagramme en barres, on associe chaque modalité avec un rectangle, sa longueur proportionnelle à l'effectif de cette modalité, on constate que la modalité la plus élevèe est "Beaucoup" qui précède tout juste "Passionnément" loin devant les dernières modalites "Un peu" et "Pas du tout".*

Pour le diagramme circulaire on associe chaque modalité avec un secteur du cercle.

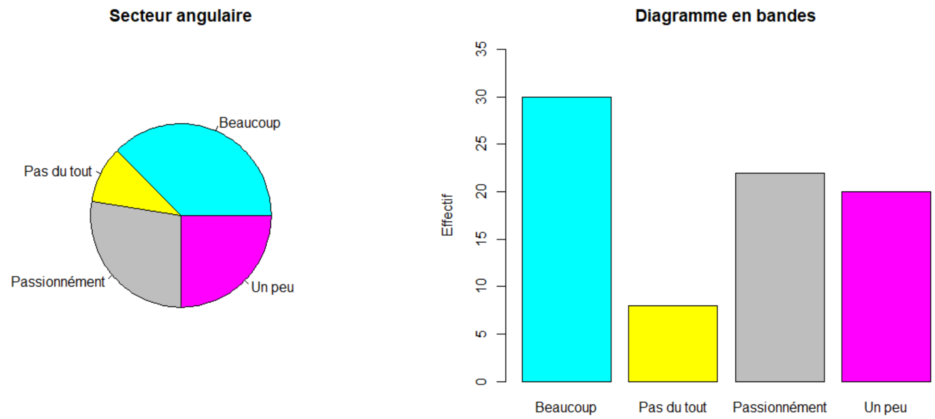

FIG. 3.1 – "secteur angulaire" et "le diagramme en bandes".

3.2 Exemples sur la variable quantitative

3.2.1 Cas d'une variable quantitative discrète

Exemple 3.2.1 *: Nous avons fait une étude sur 120 familles afin de connaître le nombre d'enfants scolarisés dans chaque famille, et les résultats sont présentés dans le tableau ci-dessous :*

Nombre d'enfants scolarisés	0	1	2	3	4	5	**Totale**
Nombre des familles	12	30	18	39	6	15	**120**

TAB. 3.2 – Tableau représente le nombre d'enfants scolarisés dans 120 familles.

1. Déterminer les paramètres caractéristiques.
2. Tracer le diagramme en bâtons et le diagramme cumulatif.

Solution 3.2.1 *:*

· *La population statistique* ***: les familles****, échantillon* ***: 120 familles****, l'unité statistique :* ***une famille,*** *la variable statistique est* ***le nombre d'enfants scolarisés par chaque famille*** *et elle est* ***quantitative discréte****.*

Lecture des données :

> X=c(0,1,2,3,4,5)   # Nombre d'enfants scolarisés.

> Y=c(12,30,18,39,6,15)   # Nombre des familles.

> D=rep(X,Y)

> T=table(D)

Les paramètres de tendance centrale :

Les paramètres	Le mode	La mediane	La moyenne
Code **R**	which(T==max(T))	median(D)	mean(D)
Les résultats	3	2.5	2.35

Les paramètres de dispersion :

Les paramètres	Code **R**	Les résultats
L'étendue	max(D)-min(D)	5
La variance	var(D)	2.195798
L'écart-type	sd(D)	1.481823
Le cofficient de variation	sd(D)/mean(D)	0.6305628
L'écart absolu moyen	mean(abs(X-mean(D)))	1.5
Intervalle interquartile	IQR(D)	2

Les quartiles : Code R $\Longrightarrow$ quantile(D)

Les résultats :

0%	25%	50%	75%	100%
0.0	1.0	2.5	3.0	5.0

Commentaire : *On remarque que : le cofficient de variation est grand* (63%), *ce qui signifie que la dispersion de la distribution est plus forte.*

Les paramètres de forme :

Nous utilisons la fonction skewness et kurtosis du Package moments.

	Code **R**	Les résultats
Coefficient d'asymétrie	skewness(D)	0.2243094
Coefficient d'aplatissement	kurtosis(D)	2.209315

Commentaire : *On voit sur ce dernier que la moyenne, la médiane et le mode sont approximativement égaux* ($\overline{X} \backsimeq M_e \backsimeq M_o$) *et la valeur du coefficient d'asymétrie prèsque nulle. Alors la distribution est symétrique.*

D'autre part, la valeur du coefficient d'aplatissement est inférieur à 3, c'est à dire la distribution des données est platykurtique.

2. La représentation graphique : Instruction R

- ***Pour tracer le diagramme en bâtons*** *: On utilise la fonction plot* (X, Y).

- ***Pour tracer le diagramme cumulatif*** *: On utilise la fonction plot (ecdf(D)).*

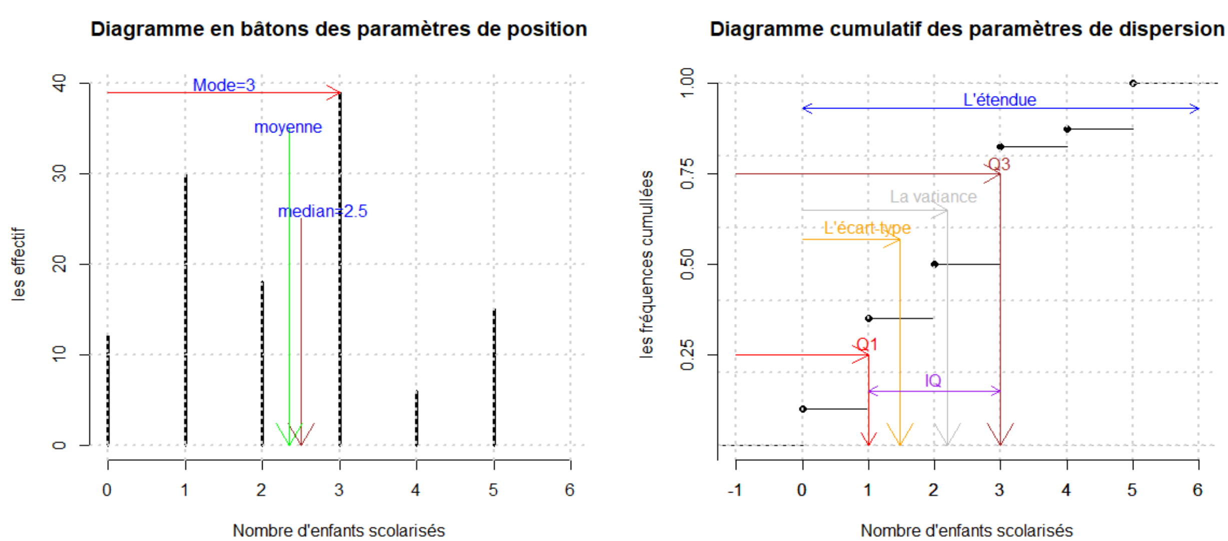

Fig. 3.2 – Les diagrammes d'une variable quantitative discrète.

Commentaire : *Dans le diagramme en bâtons : On a déterminé les paramètres de position, où le plus haut bâton est le mode, ce qui signifie que la plupart des familles ont 3 enfants scolarisés. Et on a : la médiane qui prend la valeur 2.5, ce qui signifie* *que* 50% *des familles ont un nombre d'enfants entre* $0, 1$ *et* 2, *et l'autre des familles ont plus de* 2 *enfants, et la valeur centrale de cet example est* $\overline{X} = 2.35$.

Dans la courbe cumulative : On ajoute des fléches pour clarifier les paramètres de dispersion.

3.2.2 Cas d'une variable quantitative continue

Exemple 3.2.2 *: (Effectifs groupés par classes d'amplitudes inégales)**,** temps passé devant des médias sociaux par 150 élèves pendant une certaine journée :*

Temps en minutes	$[15, 25[$	$[25, 35[$	$[35, 50[$	$[50, 65[$	$[65, 75[$	$[75, 90[$	**Total**
n_i	18	12	30	54	15	21	150

1. Calculer les f_i, ECC, FCC, n_i', f_i'.

2. Déterminer les paramètres caractéristiques.

3. Tracer l'histogramme des fréquences corrigées.

Solution 3.2.2 *:*

· *La population statistique* ***: les élèves****, échantillon* ***: 150 élèves****, l'unité statistique :* ***un élève,*** *la variable statistique est* ***le temps passé devant des médias sociaux par des élèves*** *et elle est* ***quantitative continue****.*

1. *Lecture des données :*

> Ci=c(20,30,42.5,57.5,70,82.5)   # Centre de classe.

> Z=c(18,12,30,54,15,21)   # Effectif partiel.

> K=rep(Ci,Z)

Le logiciel ***R*** *nous donne les résultats suivants :*

	a	n_i	ECC	f_i	FCC	n_i'	f_i'
[15,25[	*10*	*18*	*18*	*0.12*	*0.12*	*1.8*	*0.012000000*
[25,35[	*10*	*12*	*30*	*0.08*	*0.20*	*1.2*	*0.008000000*
[35,50[	*15*	*30*	*60*	*0.20*	*0.40*	*2.0*	*0.013333333*
[50,65[	*15*	*54*	*114*	*0.36*	*0.76*	*3.6*	*0.024000000*
[65,75[	*10*	*15*	*129*	*0.10*	*0.86*	*1.5*	*0.010000000*
[75,90[	*15*	*21*	*150*	*0.14*	*1.00*	*1.4*	*0.009333333*

2. ***Les paramètres caractéristiques :***

Les paramètres de tendance centrale :

Le mode	*La mediane*	*La moyenne*
54.32432	*54.16667*	*52.55*

Commentaire : *On voit sur ce dernier que :* $M_o \backsimeq M_{ed} > \overline{X}$*, ce qui implique la distribution est étalée à gauche.*

Les paramètres de dispersion et de forme :

E	Var	σ_X	CV
62.5	355.2408	18.84783	0.3586647
min	1^{er} *quartile*	$3^{éme}$ *quartile*	IQ
20	38.75	64.58333	25.83333
max	e_{moy}	*asymétrie*	*aplatissement*
82.5	19.58333	-0.139186	2.245167

Commentaire : *On observe que :*

- Le coefficient de variation est petit, ce qui signifie que la dispersion de la distribution est faible.

- Le coefficient d'asymétrie (skewness) est négatif, ce qui implique que la distribution est décalée à droite de la médiane, et donc une queue de distribution étalée vers la gauche.

- Le coefficient d'aplatissement est inférieur à 3, c'est à dire la distribution des données est platykurtique.

3. La représentation graphique : Instruction R .

*- **Pour tracer l'Histogramme des fréquences corrigées :** On utilise la fonction hist(K,prob=TRUE).*

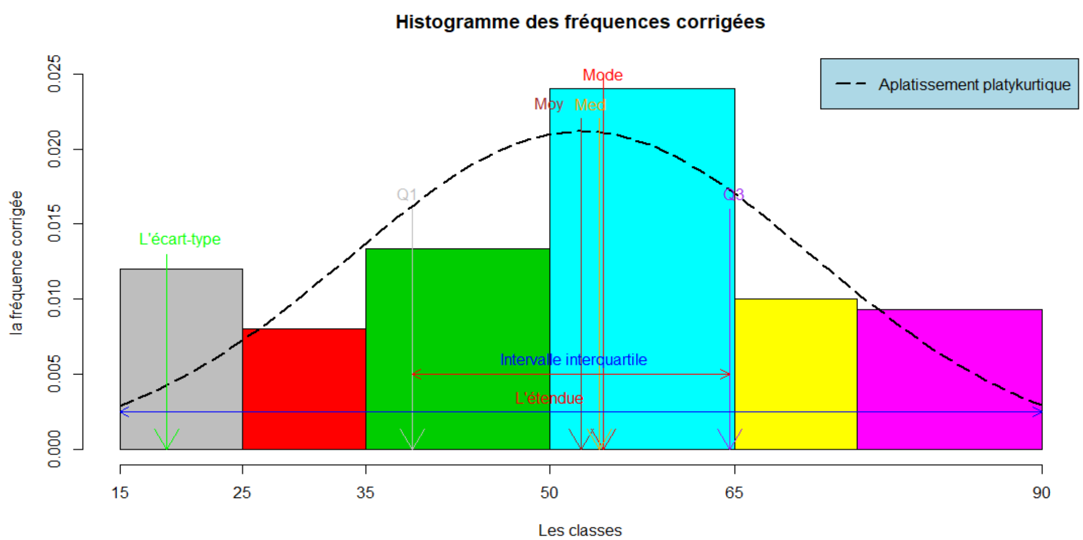

FIG. 3.3 – Histogramme des paramètres caractéristique.

Commentaire : *Dans l'histogramme, on ajoute des fléches pour clarifier les paramètres de tendance centrale et de dispersion, on remarque que la valeur* 54.16 *signifie que le nombre moyen d'élèves passent* 54.16 *minutes sur les médias sociaux, cette valeur divise l'histogramme en deux parties prèsque de même surface.*

Conclusion

Ce mémoire donne une idée générale sur la statistique descriptive univariée, en premier lieu nous avons réalisé comment obtenir des résultats précis et clairs, à partir de la représentation graphique (histogramme, camembert graphique...) de chaque variable, ensuite nous avons défini certaines caractéristiques (caractéristiques de tendance centrale, caractéristiques de dispersion, caractéristiques de forme) en donnant des exemples explicatifs, mais les résultats obtenus ne peuvent être catégoriquement généralisés ni prédire le comportement du phénomène et ses connaissances dans le présent ou le futur, ce qui nous incite à recourir aux statistiques inférentielles.

Enfin, nous illustrons notre principal résultat de ce mémoire que la statistique descriptive univariée est une méthode et technique importante sur laquelle s'appuit la science statistique.

Bibliographie

[1] Ayache, A.,Hamonier, J.Cours de Statistique Descriptive.
htttp ://math.univ-lille.fr / $\sim$ ayache/ cours-SD.pdf.

[2] Baccini, A. (2010). Statistique descriptive élémentaire.Institut de Mathématiques de Toulouse.

[3] Betteka. S., (2018 - 2019). Statistique descriptive univariée.

[4] Chekroun, A. (2017 - 2018). Statistiques descriptives et exercices.

[5] Desgraupes, B. (2018 – 2019). Statistiques Descriptives.

[6] DIOURI, M., ELMARHOUM, A.(2006). STATISTIQUE DESCRIPTIVE Cours et exercices.

[7] Hammdani, H. (2001). Statistique descriptive. OFFICE DES PUBLICATIONS UNIVERSITAIRES : 02-2001.

[8] IMMEDIATO, H. (2001). LICENCE Scientifique. Cours. Statistiques.

[9] Meghlaoui, D. (2010). Introduction à la Statistiques descriptives.

[10] SELLAM, M.,(2018). Statistique descriptive univariée.

[11] Tillé, Y. (2010). Résumé du Cours de Statistique Descriptive.

[12] Torrés. O. (2007). Cours de statistique descriptive. 1. Analyse univariée.

3.2.3 Annexe A : Logiciel R

3.3 Qu'est-ce-que le langage R ?

• Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• R a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

Symbole		**Signification**
Ω	:	Population l'ensemble statistique.
ω	:	Individu ou Unité statistique.
X	:	Caractère ou variable statistique.
x_i	:	Modalités du caractère X.
n	:	Effectif total.
n_i	:	Effectif partiel.
f_i	:	Fréquence partielle.
$ECC\ N_i$	:	Effectifs cumulés croissants.
$ECD\ N_i^{'}$	:	Effectif cumulé décroissante.
$FCC\ F_i$	:	Fréquences cumulées croissantes.
$FCD\ F_i^{'}$	:	Fréquences cumulées décroissantes.
$f_i^{'}$	:	La fréquence corrigée.
$n_i^{'}$	:	L'effectif corrigé.
$card(\Omega)$	:	Le cardinal : nombre d'éléments de l'ensemble Ω.

b_i, b_{i+1}	:	Les bornes d'une classe.
a_i	:	L'amplitude d'une classe.
c_i	:	Le centre d'une classe.
$\sum_{i=1}^{i=k}$	:	La somme pour i variant de 1 à k.
M_o	:	Le mode.
M_e	:	La médiane.
$\overline{X}$	:	La moyenne arithemétique de X.
E	:	L'étendue.
$Q_{(1,2,3)}$	:	Les quartiles d'ordre $(1, 2, 3)$.
$Var(X)$	:	La variance de X.
σ_X	:	L'écart-type de X.
$CV(X)$	:	Coefficient de variation de X.
m_a^p	:	Moment d'ordre p
Y	:	Le coefficient de Yule.
P	:	Le coefficient de Pearson.
F	:	Le coefficient de Fisher.
β_2	:	Coefficient d'aplatissement de Pearson.
F_2	:	Coefficient d'aplatissement de Fisher.

Résumé

La statistique descriptive univariée est l'étude de données associées d'une seule variable, que celle-ci soit d'une variable qualitative ou quantitative, qui s'intéresse à résumer le phénomène étudié avec deux techniques:
La représentation graphique: donne une forme globale sur la distribution des données et simplifie les résultats, et les paramètres caractéristique (position, dispersion, forme): qui donne des interprétations de résultats obtenues.

Les Mots clés

Statistique descriptive, variable statistique, quantitative, qualitative, la représentation graphique, les paramètres caractéristique : position, dispersion, forme.

Abstract

Univariate descriptive statistics is the study of associated data of a single statistical variable, whether that variable is a qualitative or a quantitative variable. Which is interested in summarizing the phenomenon studied with two techniques:
The graphical representations: gives a global shape on the distribution of the data and simplifies the results, and the characteristics parameters (the position, the dispersion, the shape): which gives interpretations of the results obtained.

Key words

Descriptive statistics, statistical variable, quantitative, qualitative, the graphical representations, the characteristics parameters: the position, the dispersion, the shape)

الملخص

الإحصاء الوصفي أحادي المتغير هو دراسة البيانات المرتبطة بمتغير واحد سواء كان متغيرا نوعيا أو كميا، حيث يهتم بتلخيص الظاهرة المدروسة بتقنيتين:
التمثيل البياني: يعطي شكلا شاملا لتوزيع البيانات ويبسط النتائج، والمعلمات المميزة:(الموضع، التشتت، الشكل): التي تعطي تفسيرات للنتائج التي تم الحصول عليها.

الكلمات مفتاحية

الإحصاء الوصفي، متغير إحصائي، كمي، نوعي، التمثيل البياني، المعلمات المميزة: (الموضع، التشتت، الشكل).