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

SELLAM Mounira

Titre :

La statistique descriptive univariée

Membres du Comité d'Examen :

Dr.	**CHERFAOUI Mouloud**	UMKB	Président
Dr.	**SAYAH Abdallah**	UMKB	Encadreur
Pr.	**MERAGHNI Djamel**	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce modeste travail à :

A l'homme de ma vie, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir, que Allah te garde dans son vaste paradis, à toi mon père.

A lumière de mes jours, la source de mes efforts, la flamme de mon Cœur, ma vie et mon bonheur ; maman que j'adore.

A mon encadreur, Monsieur «Sayah Abdallah».

A mes chers sœurs et frères, pour leurs encouragements permanents, et leur soutien moral.

A mon ami préféré Zouaoui Nour Elhouda, à tous mes amis.

A tous ceux qui m'aiment et que j'aime.

REMERCIEMENTS

Je tiens à remercier premièrement, Allah qui m'a donné le pouvoir, la volonté, et le courage et les bonnes chances pour réussir. Je remercie vivement mon encadreur, Monsieur «Sayah Abdallah», ses précieux conseils et son aide durant toute la période du travail.

J'aimerais présenter mes remerciements et ma gratitude aux membres du jury, Monsieur Le Dr CHERFAOUI Mouloud et Madame BENBRIKA Ghozlane, d'avoir examiner et évaluer mon travail

Je n'oublie pas de remercier tous les enseignants de département de Mathématiques.

J'adresse un grand merci à mes parents, à mes sœurs et frères, à tous ma famille, à mon ami préféré Zouaoui Nour Elhouda.

A toute personne qui a participé de près ou de loin à l'exécution de ce modeste travail.

Merci A tout pour tout.

Table des matières

Table des figures

Liste des tableaux

Introduction

La statistique est une méthode scientifique qui sert à collecter, analyser et traiter un ensemble de données communément appelé en statistique “ensemble de variables”. Principalement, il existe deux groupes de variables : les variables quantitatives (discrètes et continues) et les variables qualitatives (ordinales et nominales). La première étape d'une analyse de données est de connaître ces différents types de variables, pour choisir les différentes méthodes statistiques correspondantes.

Elle décrit le phénomène à partir des représentations graphiques (histogramme, diagramme en bâtons, diagramme circulaire, et les diagrammes cumulatifs...), et les paramètres caractéristiques (effectif, fréquence, mode, moyenne, variance, écart-type, étendue...) qui permettent de résumer en quelques chiffres les distributions des variables qui constituent. Dans ce mémoire, on présente l'un des branches de la statistique qui est la statistique descriptive univariée.

Ce travail se partage en deux chapitres :

Le premier chapitre est consacré à une introduction générale sur les principaux concepts qu'ils conviennent de connaître les notions de base, et comment les organiser dans un tableau statistique et les tracer graphiquement.

Dans le deuxième chapitre, on offre une présentation détaillée sur les indicateurs caractéristiques (les indicateurs de position, les indicateurs de dispersion et les indicateurs de forme).

Enfin, on va appliquer tout ce qu'on a vu dans les deux chapitres, avec un exemple général.

Chapitre 1

Généralités

La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation graphique, avec des méthodes de classement tels que, les tableaux statistiques et les graphiques.
Dans ce chapitre, on s'intéresse d'introduire les notions de base de la statistique descriptive (l'effectif, la fréquence, la fréquence cumulée...), et de traiter les données d'une série statistique à partir d'organisation de ces données sous forme des tableaux statistiques, et les présenter graphiquement.

1.1 Définitons fondamentales

Définition 1.1.1 *(Statistique)*
On appelle statistique l'ensemble des méthodes scientifiques qui permettent de recueillir, organiser, classer et présenter des informations statistiques qualitatives ou quantitatives, pour tirer des conclusions sur la population étudiée.

Son objectif est d'extraire des informations pertinentes d'une liste de nombres difficile à interpréter par une simple lecture.

1.1.1 La statistique descriptive

Le principe de cette méthode statistique est la description des données étudiées à l'aide de moyens appropriés, qu'ils correspondent à des valeurs calculées (moyenne, médiane, écart-type, quartiles.....) ou à des représentations graphiques (Histogramme, camembert, ...) et comme un exemple l'étude de la distribution des salaires dans une entreprise. L'objectif de la statistique descriptive est de résumer l'échantillon par deux moyens : l'approche numérique et l'approche 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 taille, le poids.....).
- **La statistique descriptive multivariée :** Est l'étude de la relation qui peut exister entre deux ou plusieurs variables, que l'on traite avec des méthodes comme l'analyse factorielle. (Par exemple la relation entre la taille et le poids. . . .etc.).

1.1.2 Vocabulaires statistiques

- **Population :** La population est l'ensemble d'éléments homogènes sur lequel on effectue une analyse statistique, elle est notée Ω. Par exemple : les étudiants d'une classe, les véhicules automobiles immatriculés en Algérie, ...etc.
- **Individu (ou unité statistique) :** Un individu est un élément de la population. Il est noté ω.
- **Echantillon :** L'échantillon est un sous ensemble de la population statistique.
- **Les modalités :** Les modalités sont les différentes situations x_i possibles du caractère (ou les valeurs possibles de X), où chaque caractère possède deux ou plusieurs modalités.

1.1.3 Un caractère (ou une variable)

Définition 1.1.2 *Une variable X est un moyen de décrire chacun des individus de la population étudiée, par exemple (âge, salaire, sexe,...). Elle a deux types différents suivants :*

Les types de variable :

Une variable quantitative : Une variable est dite quantitative, si l'ensemble des observations est un ensemble des nombres numériques qu'ils peuvent être ordonnés, et on distingue deux types de variables quantitatives :

1. *Une variable quantitative discrète :* Elle ne prend que des valeurs entiers : 0, 1, 2, 3....etc. Elle peut être représentée par un nombre fini de valeurs, par exemple : le nombre d'enfants par famille.
2. *Une variable quantitative continue :* Elle est dite continue, lorsque ces modalités ne sont pas des valeurs précises, mais des intervalles $[a; b]$ de nombre réels. Par exemple : le poids, la taille, l'âge....etc.

Une variable qualitative : Une variable est dite qualitative lorsque les modalités d'une variable sont des catégories que l'on désigne par des noms, qu'elles ne peuvent pas s'exprimer par des nombres. Par exemple : la couleur de peau est une variable a pour modalité : blanc, noir, jaune, rouge,...etc. Et elle a deux catégories :

1. *Une variable qualitative ordinale :* Elle est dite ordinale quand les modalités peuvent être naturellement ordonnées, par exemple : niveau d'études, classe sociale, grade,.....etc.
2. *Une variable qualitative nominale :* Elle est dite nominale lorsque ses modalités ne peuvent être classées de façon naturelle par exemple : la variable couleur des yeux et la variable sexe.....etc.

1.1.4 Effectifs, fréquences, fréquences cumulées

Effectif :

Définition 1.1.3 *L'effectif d'une modalité x_i est le nombre de fois, où cette modalité apparait dans la série statistique, on le note n_i. On a :*

$$n_i = card\{\omega \epsilon \Omega, \quad X(\omega) = x_i\}. \tag{1.1}$$

Effectif total :

Définition 1.1.4 *L'effectif total est le nombre total d'observations, c'est aussi la somme des effectifs de chaque valeur ; on le note n :*

$$n = \sum_{i=1}^{p} n_i = card(\Omega). \tag{1.2}$$

Fréquence (fréquence relative) :

Définition 1.1.5 *La fréquence d'une modalité est le quotient de son effectif par l'effectif total, elle est notée f_i :*

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

Remarque 1.1.1 *La valeur de la fréquence est toujours comprise entre 0 et 1.*
On peut remplacer f_i par $f_i \times 100$, si on exprime les fréquences en pourcentage.
$\sum_{i=1}^{p} f_i = 1$, ou $\sum_{i=1}^{p} f_i(\%) = 100$, le cas des fréquences en pourcentage.

Effectif et fréquence cumulés :

Définition 1.1.6 *L'effectif cumulé croissant (la fréquence cumulée croissante) d'une modalité de rang i est la somme de tous les n_i (de toutes les f_i) jusqu'au rang i compris. Ils*

sont notés respectivement ECC et FCC [1] *:*

$$ECC = \sum_{j=1}^{i} n_j;\ FCC = \sum_{j=1}^{i} f_j. \tag{1.4}$$

L'effectif cumulé décroissant (la fréquence cumulée décroissante) d'une modalité de rang i est la somme de tous les n_i (de toutes les f_i), à partir de la dernière valeur jusqu'au rang i compris. Ils sont notés respectivement ECD et FCD :

$$ECD = \sum_{j=i}^{p} n_j;\ FCD = \sum_{j=i}^{p} f_j. \tag{1.5}$$

1.2 Représentation des données

Il existe plusieurs méthodes de la description statistique : la présentation brute des données, des représentations par des tableaux statistiques numériques, des représentations graphiques, celui nous permet de visualiser rapidement les informations, et d'avoir une vue plus globale du phénomène étudié.

1.2.1 Série statistique

Définition 1.2.1 *Une série statistique est l'ensemble des différentes données associées à un certain nombre d'individus.*

1.2.2 Tableau statistique

Définition 1.2.2 *Un tableau statistique est un moyen d'organiser, classifier et ranger par ordre croissant (ou décroisant) les données brutes de la série statistique, pour les bien représenter.*

[1] Les effectifs cumulés croissants et les fréquences cumulées croissantes notées N_i et F_i respectivement.
– ECC représente le nombre d'observations inférieures ou égales à x_i, et FCC leur fréquence. On note que $N_p = n$ et $F_p = 1$.

On crée un tableau statistique synthétique, où les observations qui ont même modalité sont regroupées comme des valeurs numériques pour les variables quantitatives discrètes, et comme des intervalles pour les variables continues. On s'intéresse à deux types de tableaux qui dépendent du type de variables.

Caractère qualitatif et quantitatif discret :

Les modalités $\mathbf{x}_i$	**Effectifs $\mathbf{n}_i$**	**ECC $\mathbf{N}_i$**	**Fréquence $\mathbf{f}_i$**	**FCC $\mathbf{F}_i$**
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_p	$N_p = n$	f_p	$F_p = 1$
Total	n		1	

TAB. 1.1 – Tableau statistique d'un caractère qualitatif et quantitatif discret

Caractère quantitatif continu :

Les classes	**Centres $\mathbf{c}_i$**	**l'amplitude $\mathbf{a}_i$**	**Effectifs $\mathbf{n}_i$**	**ECC $\mathbf{N}_i$**	**Fréquence $\mathbf{f}_i$**	**FCC $\mathbf{F}_i$**
$[\mathbf{b}_1; \mathbf{b}_2[$	c_1	a_1	n_1	N_1	f_1	F_1
$[\mathbf{b}_2; \mathbf{b}_3[$	c_2	a_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[\mathbf{b}_p; \mathbf{b}_{p+1}[$	c_p	a_p	n_p	$N_p = n$	f_p	$F_p = 1$
Total			n		1	

TAB. 1.2 – Tableau statistique regroupé par classes

Remarque 1.2.1 $\mathbf{b}_i$ *et* $\mathbf{b}_{i+1}$ *sont les bornes d'une classe.*
Le centre d'une classe est : $\mathbf{c}_i = \dfrac{\mathbf{b}_i + \mathbf{b}_{i+1}}{2}$.
L'amplitude d'une classe est : $\mathbf{a}_i = \mathbf{b}_{i+1} - \mathbf{b}_i$.

1.3 Représentation graphique

En général, la représentation graphique des données relatives à un caractère unique est une synthèse de l'information qui fait apparaitre la forme globale de la distribution des données. La nature de graphique dépend du type de variables.

1.3.1 Cas d'une variable quantitative

Pour les variables quantitatives, il existe deux types de représentation graphique qui sont :

Les diagrammes différentiels :

Cas d'une variable quantitative discrète :

Le diagramme en bâtons : Ce diagramme comporte deux axes, un axe horizontal qui représente les valeurs de la variable, et un axe vertical qui représente les effectifs ou les fréquences, à chaque valeur on associe un segment (bâton) dont sa hauteur est proportionnelle à l'effectif ou à la fréquence de cette modalité.

Cas d'une variable quantitative continue :

L'histogramme : L'histogramme est un graphique qui représente des rectangles ayant pour base les bornes des classes, (où l'amplitude de ces classes est la largeur des rectangles) et leurs hauteur est proportionnelle à la densité d'effectif (ou la densité de fréquence) définie comme suit :

$$f_i^{'} = \frac{f_i}{a_i} \quad \text{et,} \quad n_i^{'} = \frac{n_i}{a_i}, \tag{1.6}$$

avec $f_i^{'}$ est la fréquence corrigée (**La densité de fréquence)**, $n_i^{'}$ est l'effectif corrigé (**La densité d'effectif)**.

La surface du rectangle est égale à l'effectif ou à la fréquence d'une modalité.

Remarque 1.3.1 *La surface de l'histogramme est égale à l'effectif total n si on travaille avec les effectifs, et elle est égale à* 1 *si on travaille avec les fréquences.*

Les diagrammes cumulatifs :

Les diagrammes cumulatifs permettent de visualiser l'évolution des fréquences cumulées ou les effectifs cumulés croissants ou décroissants, ils sont obtenus à partir de la fonction de répartition empirique.[2]

Exemple 1.3.1 *(Pour une variable quantitative discrète)*
Un quartier est composé de 45 ménages, la variable ici est le nombre de personnes par ménages, d'après le calcul des effectifs, les effectifs cumulés, les fréquences, les fréquences cumulées, on construit le tableau statistique suivant :

Nombre de pessonnes par ménages	$\mathbf{n}_i$	$\mathbf{N}_i$	**ECD**	$\mathbf{f}_i$	$\mathbf{F}_i$	**FCD**
1	3	3	45	0.07	0.07	1
2	8	11	42	0.18	0.25	0.93
3	14	25	34	0.31	0.56	0.75
4	9	34	20	0.2	0.76	0.44
5	6	40	11	0.13	0.89	0.24
6	3	43	5	0.07	0.96	0.11
8	2	45	2	0.04	1	0.04
Total	45			1		

TAB. 1.3 – Tableau de nombre des personnes par ménages

Par exemple, si on veut expliquer comment obtenir l'effectif cumulé croissant 25 de la modalité $x_3 = 3$, qui signifie que 25 ménages ont 3 personnes au maximum. Il suffit d'ajouter

[2]La fonction de répartition empirique :
– $F : \mathbb{R} \longmapsto [0;1]$
$X \longmapsto F(x) = P(X \leq x)$.

l'effectif cumulé croissant précédent de la modalité $x_2 = 2$ et l'effectif correspondant à $x_3 = 3$. De même manière on complète les autres ECC et FCC.

Pour les ECD et FCD, on obtient l'effectif cumulé décroissant 20 de la modalité $x_4 = 4$, qui signifie que 20 ménages ont 4 personnes au minimum, il suffit d'ajouter l'effectif correspondant à $x_4 = 4$ et les effectifs qui lui suivant. On obtient les diagrammes suivants :

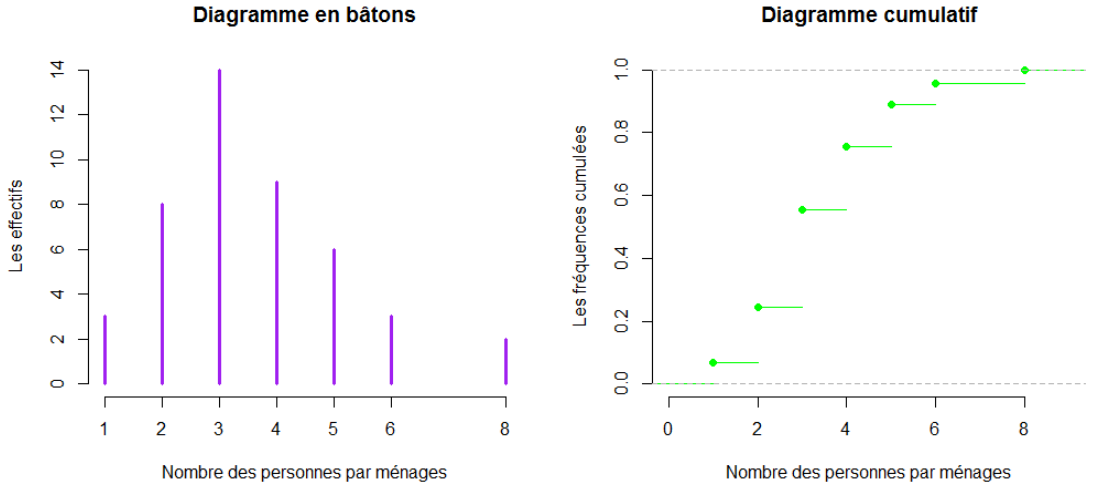

FIG. 1.1 – Les diagrammes d'une variable quantitative discrète

Pour le diagramme en bâtons, on associe chaque modalité avec un segment, dont sa hauteur proportionnelle à l'effectif de cette modalité.

Pour le diagramme cumulatif, son principe est de tracer des points d'abscisse x_i et d'ordonnée F_i, on complète la représentation graphique par des paliers, en obtenant une courbe en escalier, on passe de niveau 0 au niveau 0.07 avec une hauteur égale à $f_1 = 0.07$, et de niveau 0.07 au niveau 0.25 d'une hauteur égale à 0.18, et ainsi de suite.

Exemple 1.3.2 *(D'une variable quantitative continue)*

On mesure la taille en centimètres de 50 *élèves d'une classe. On obtient le tableau suivant :*

Classes	$\mathbf{n}_i$	$\mathbf{c}_i$	$\mathbf{a}_i$	**ECC**	**ECD**	$\mathbf{f}_i$	**FCC**	**FCD**	$\mathbf{n}_i' = \frac{\mathbf{n}_i}{\mathbf{a}_i}$	$\mathbf{f}_i' = \frac{\mathbf{f}_i}{\mathbf{a}_i}$
[150; 160[	15	155	10	15	50	0.3	0.3	1	1.5	0.03
[160; 165[	7	162.5	5	22	35	0.14	0.44	0.7	1.4	0.028
[165; 170[	11	167.5	5	33	28	0.22	0.66	0.56	2.2	0.044
[170; 175[	14	172.5	5	47	17	0.28	0.94	0.34	2.8	0.056
[175; 185[	3	180	10	50	3	0.06	1	0.06	0.3	0.006
Total	50					1				

TAB. 1.4 – Tableau du taille en centimètres des élèves

Comme les classes de cet exemple ne sont pas d'égale amplitude, on calcule la densité d'effectif, dont la hauteur de chaque rectangle proportionnelle à la densité d'effectif correspondant, comme on va voir dans la figure(1.2).

Pour la courbe cumulative, son principe est de tracer des points d'abscisse b_{i+1} et d'ordonnée F_i, sauf le premier point qui a comme abscisse 150 et d'ordonnée 0, on complète la représentation graphique en rejoignant les points par des segments pour obtenir la courbe cumulative continue. On construit les diagrammes suivants :

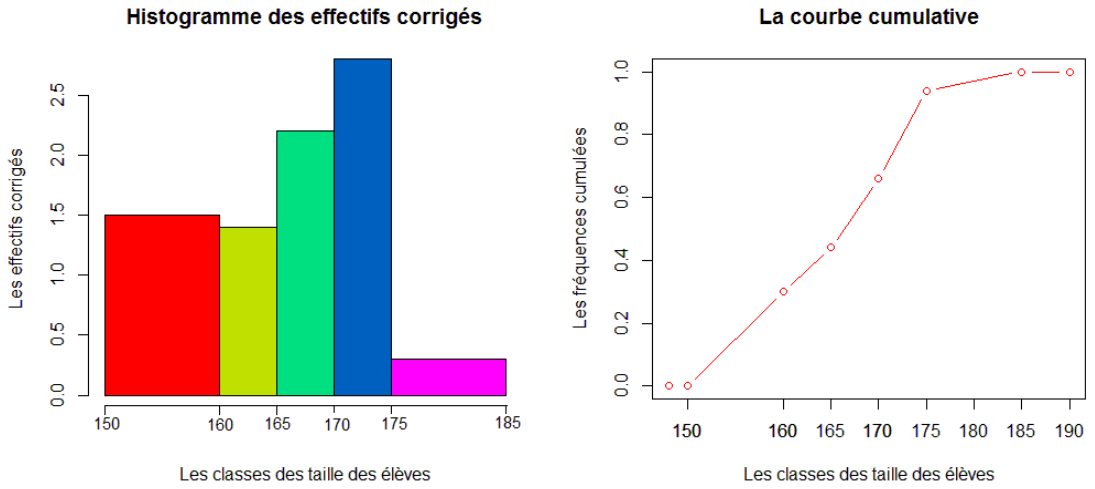

FIG. 1.2 – Les diagrammes d'une variable quantitative continue

1.3.2 Cas d'une variable qualitative

Lorsque le caractère est qualitatif, on utilise le tableau de fréquence pour construire les graphiques, qui permettent de représenter la série statistique.

Diagramme circulaire (diagramme en secteurs ou camembert) :

Le principe du graphe consiste à diviser un cercle ou un disque en secteurs. Chaque secteur représente une modalité sa surface est proportionnelle à la fréquence de cette modalité. Pour une modalité donnée x_i, l'effectif n_i, l'angle au centre α_i correspondant est (en degré) :

$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360. \tag{1.7}$$

L'effectif total est représenté par le disque de figure.(1.3).

Diagramme en barres :

Le diagramme en barres est un ensemble de rectangles de même largeur, séparés par un espace, qu'ils sont placés sur un axe horizontal (cette droite n'orient pas car les modalités ne sont pas numériques et elles n'ont pas de relation d'ordre), chaque rectangle représentant une modalité, sa hauteur est proportionnelle à l'effectif ou à la fréquence de la modalité.

Exemple 1.3.3 *On s'intéresse à une série statistique du variable " état civil " sur* 20 *personnes. On obtient le tableau statistique suivant :*

$\mathbf{x}_i$	**Effactif** $\mathbf{n}_i$	**Fréquence** $\mathbf{f}_i$	$\mathbf{f}_i$ **en pourcentage** (%)
Célibataire	8	0.4	40
Marié(e)	6	0.3	30
Veuf(ve)	3	0.15	15
Divorcé(e)	3	0.15	15
Total	20	1	100

TAB. 1.5 – Tableau d'état civil

Dans cet exemple, on utilise la colonne de fréquences en pourcentage pour le diagramme en secteurs, où chaque secteur présente une modalité qui a comme surface la fréquence ($\mathbf{f}_i(\%)$) correspondant. Et pour le diagramme en barres, on associe chaque modalité avec un rectangle, sa longueur proportionnelle à l'effectif de cette modalité. Les modalités ont été positionnées dans l'ordre alphabétique, on constate que la modalité la plus fréquente est l'état «Célibataire», qui précède tout juste «Marié(e)», loin devant les deux dernières modalités «Divorcé(e)» et «Veuf(ve)», comme le montre les deux diagrammes suivants :

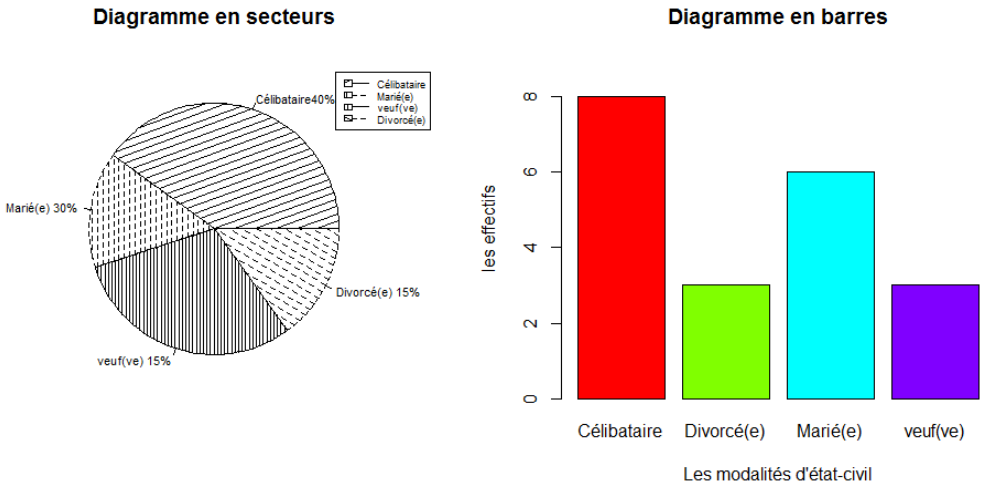

FIG. 1.3 – Les diagramme d'une variable qualitative

Chapitre 2

Statistique descriptive univariée

La statistique descriptive univariée consiste en la description de chacun des caractères statistiques, un par un, et non des liens éventuels existant entre eux, elle sert à présenter les données observées d'une variable, sous forme de tableaux et de graphiques, et les résumer numériquement avec des indicateurs statistiques, tous ces moyens sont des outils et des méthodes de la statistique descriptive univariée. Ces méthodes donnent une information simple à manipuler sur la série statistique.

Dans le premier chapitre, on a vu le traitement graphique des données, dans ce chapitre, on s'intéresse à étudier les paramètres caractéristiques d'une seule variable.

2.1 Paramètres caractéristiques

L'objectif d'une étude statistique est aussi de résumer et visualiser les données par des paramètres ou des indicateurs caractéristiques, qu'ils sont séparés par trois types : les paramètres de position, les paramètres de dispersion et les paramètres de forme.

2.1.1 Paramètres de position

Les paramètres de position donnent une idée sur la position des données, ils permettent à indiquer autour de quelle valeur centrale se situent ces données.

Le mode :

Définition 2.1.1 *Le mode est la modalité qui a le plus grand effectif (ou la plus grande fréquence). Pour les variables quantitatives continues, on parle de la classe modale qui constitue le mode de la distribution. Si les classes sont d'amplitude égale, la classe modale est la classe présentant l'effectif ou la fréquence les plus élevés, si les classes sont d'amplitude différente, alors la classe modale est la densité d'effectif (ou la densité de fréquence) les plus élevés(1.6). Il est noté : M_o ou x_M. Et on le calcule numériquement comme suit :*

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

où, b_i est la borne inferieure de la classe modale,

- b_{i+1} est la borne supérieure de la classe modale,

- n'_{i+1} est la densité d'effectif la plus élevée,

- n'_i est la densité d'effectif précédente,

- n'_{i+2} est la densité d'effectif suivante.

Graphiquement, et sur l'histogramme, la classe modale $[b_i; b_{i+1}[$ est associée au rectangle le plus haut.

Remarque 2.1.1 *Le mode se calcule pour tous les types de variables.*

Le mode n'est pas nécessairement unique.

Il peut exister des distributions sans mode. Ce sont des distributions uniformes dont toutes les modalités ont la même fréquence ou même effectif.[5]

La médiane :

Définition 2.1.2 *Si les valeurs étant rangées par ordre croissant (ou décroissant), la médiane, désignée par M_e ou $x_{\frac{1}{2}}$, est la valeur centrale de la série statistique, qui partage cette série des observations en deux ensembles d'effectif égaux.*

– **Cas d'un caractère quantitatif discret :**

- Si n est impair, la médiane est la valeur de rang $\frac{n+1}{2}$, qui est située au milieu de la série statistique notée : $x_{(\frac{n+1}{2})}$.

$$M_e = x_{(\frac{n+1}{2})}. \tag{2.2}$$

- Si n est pair, la médiane est la moyenne de deux valeurs centrales de rang $\frac{n}{2}$ et $\frac{n}{2}+1$, notées : $x_{(\frac{n}{2})}$ et $x_{(\frac{n}{2}+1)}$.

$$M_e = \frac{1}{2} \times \left\{ x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right\}. \tag{2.3}$$

Pour cela, on commence à calculer les effectifs cumulés croissants, où la médiane est la valeur qui associe un effectif cumulé croissant supérieur ou égal au rang de la médiane (ou à la fréquence cumulée croissante égale à $\frac{1}{2}$).

– **Cas d'un caractère quantitatif continu :**

Premièrement, on cherche l'intervalle médian, de la même manière que le cas d'une variable discrète (précédente), puis, on précise la valeur de la médiane avec la méthode de l'interpolation linéaire suivante :

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

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

où, b_i est la borne inferieure de la classe médiane,

- b_{i+1} est la borne supérieure de la classe médiane,
- $\frac{n}{2}$ la moitié d'effectif total,
- N_{i+1} l'effectif cumulé croisant de la classe médiane,
- N_i l'effectif cumulé croisant précédent.

Remarque 2.1.2 *La médiane se calcule pour tous les types de variables, sauf le cas d'une variable qualitative nominale.*

L'unité de la médiane est celle de la variable.

La moyenne arithmétique :

La moyenne ne peut être définie que sur une variable quantitative.

Définition 2.1.3 *La moyenne est le rapport de la somme des valeurs observées divisées par l'effectif total, elle est notée* $\overline{X}$

– La moyenne est dite simple, lorsque chaque modalité (x_i) de la variable a un effectif (n_i).

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{p} x_i, \text{ où, } n: \text{ est l'effectif total.} \tag{2.5}$$

– Elle est dite pondérée, quand pour chaque modalité (x_i) en associant un effectif (n_i) supérieur ou égal à 1.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i. \tag{2.6}$$

- Si on travaille avec les fréquences :

$$\overline{X} = \sum_{i=1}^{p} f_i x_i, \text{ où, } f_i: \text{ sont les fréquences.} \tag{2.7}$$

Le calcul de la moyenne arithmétique :

– **Pour les variables discrètes :**

Le calcul de la moyenne arithmétique est simple, il suffit d'ajouter une colonne dans le tableau statistique contenant les produits des valeurs de la variable et celles des effectifs ou des fréquences, ensuite on somme les quantités obtenues.

– **Pour les variables continues :**

Dans ce cas, les modalités sont des classes, donc pour obtenir la moyenne, on utilise les centres des classes :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{p} n_i c_i, \text{ où, } c_i: \text{les centres des classes.} \tag{2.8}$$

Propriété 2.1.1 Propriété de linéarité de l'opérateur moyenne *:*

Si on définit une nouvelle variable Z : telle que : $z_i = ax_i + b$ *; où, a, b sont des constantes réelles. Alors* $\overline{Z} = a\overline{X} + b$*.[4]*

Preuve. On a :

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{p} n_i z_i = \frac{1}{n}\sum_{i=1}^{p} n_i(ax_i + b) \tag{2.9}$$

$$= \frac{1}{n}\sum_{i=1}^{p} an_i x_i + \frac{1}{n}\sum_{i=1}^{p} n_i b. \tag{2.10}$$

Comme a et b sont des constantes, elles ne dépendent pas du signe de sommation ($\sum$), donc on peut les faire sortir. On a alors :

$$\overline{Z} = \frac{a}{n}\sum_{i=1}^{p} n_i x_i + \frac{b}{n}\sum_{i=1}^{p} n_i, \tag{2.11}$$

or on sait que $\sum_{i=1}^{p} n_i = n$, d'où

$$\overline{Z} = \frac{a}{n}\sum_{i=1}^{p} n_i x_i + b \times \frac{n}{n} \tag{2.12}$$

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

■

Propriété 2.1.2 La moyenne des écarts à la moyenne est nulle :

$$\sum_{i=1}^{p} n_i(x_i - \overline{X}) = 0. \tag{2.13}$$

Preuve.

$$\sum_{i=1}^{p} n_i(x_i - \overline{X}) = \sum_{i=1}^{p} n_i x_i - \sum_{i=1}^{p} n_i\overline{X} = n_1(x_1 - \overline{X}) + n_2(x_2 - \overline{X}) + \ldots + n_i(x_i - \overline{X}) + \ldots + n_p(x_p - \overline{X}), \tag{2.14}$$

comme $\overline{X}$ se répète n fois, on a alors :

$$\sum_{i=1}^{p} n_i(x_i - \overline{X}) = \sum_{i=1}^{p} n_i x_i - n\overline{X}, \tag{2.15}$$

et comme $\sum_{i=1}^{p} n_i x_i = n\overline{X}$, on a alors :

$$\sum_{i=1}^{p} n_i(x_i - \overline{X}) = n\overline{X} - n\overline{X} = 0, \tag{2.16}$$

et l'égalité est vérifiée[4] ■

Exemple 2.1.1 *Pour 72 élèves des classes différentes, on connait les résultats d'un examen d'anglais, on obtient le tableau suivant :*

Les notes $\mathbf{x}_i$	$\mathbf{n}_i$	$\mathbf{c}_i$	$\mathbf{a}_i$	$\mathbf{f}_i$	$\mathbf{N}_i$	$\mathbf{n}'_i$
$[0;4[$	4	2	4	0.055	4	1
$[4;8[$	13	6	4	0.18	17	3.25
$[8;10[$	10	9	2	0.138	27	5
$[10;12[$	15	11	2	0.208	42	7.5
$[12;16[$	21	14	4	0.29	63	5.25
$[16;20[$	9	18	4	0.125	72	2.25
Total	72			1		

TAB. 2.1 – Tableau des notes des élèves

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

1. *Le mode :*

   - D'abord, on cherche la classe modale correspond à l'effectif corrigé le plus grand $n'_i = 7.5$, (comme les classes n'ont pas d'amplitude égal), elle est égale à $[10;12[$.

- Puis, on précise la valeur de mode :

$$\begin{aligned} M_o &= b_i + (b_{i+1} - b_i) \times \frac{(n'_{i+1} - n'_i)}{(n'_{i+1} - n'_i) + (n'_{i+1} - n'_{i+2})} \\ &= 10 + (12 - 10) \times \frac{(7.5 - 5)}{(7.5 - 5) + (7.5 - 5.25)} \\ M_o &= 11.05. \end{aligned}$$

2. *La médiane :*

- On a $n = 72$ (pair), dont $\frac{n}{2} = 36$, alors la classe médiane est la classe qui contient la $36^{\text{ième}}$ valeur et la $37^{\text{ième}}$ valeur, d'après la colonne N_i du tableau, on cherche l'intervalle qui a un ECC $N_i \geq 37$, on trouve que la classe médiane $= [10; 12[$

- La valeur prise est :

$$\begin{aligned} M_e &= b_i + (b_{i+1} - x_i) \times \frac{\frac{n}{2} - N_i}{N_{i+1} - N_i} \\ &= 10 + (12 - 10) \times \frac{36 - 27}{42 - 27} \\ M_e &= 11.2. \end{aligned}$$

3. *La moyenne :*

$$\begin{aligned} \overline{X} &= \frac{1}{n} \sum_{i=1}^{p} n_i c_i \\ &= \frac{1}{72} \times [(2 \times 4) + (13 \times 6) + (10 \times 9) + (15 \times 11) + (21 \times 14) + (9 \times 18)] \\ \overline{X} &= 11.069. \end{aligned}$$

- On observe que :

$M_o \simeq M_e \simeq \overline{X}$. Qui signifie que la distribution est prèsque symétrique.

- D'après l'histogramme ci-dessous, on remarque que la valeur $x_i = 11.1$ est une valeur centrale signifie que la moitié des élèves ont des notes inférieures ou égales à $x_i = 11.1$, et l'autre moitié des élèves ont des notes supérieures à $x_i = 11.1$, cette valeur partage l'histogramme en deux parties prèsque de même surface, qui signifie que la distribution est prèsque symétrique.

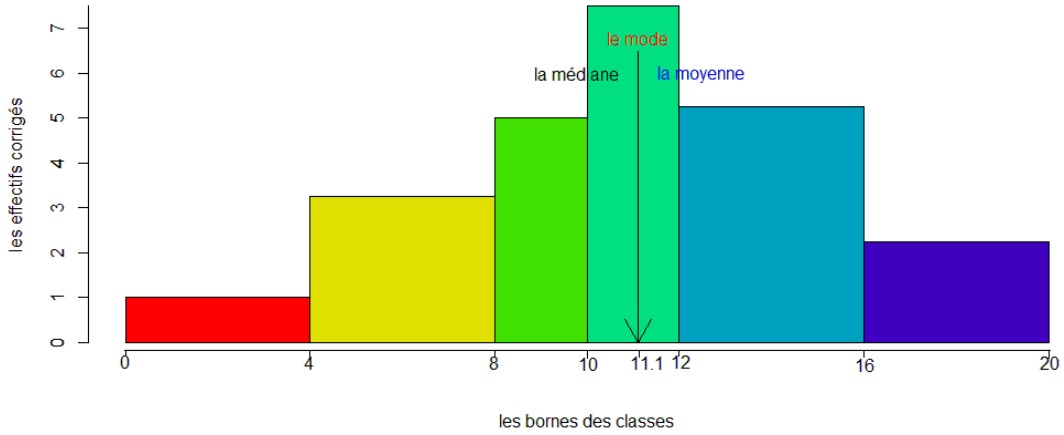

FIG. 2.1 – Histogramme des effectifs corrigés représente les paramètres de position

2.1.2 Paramètres de dispersion

Ces paramètres permettent de mesurer la variabilité (la dispersion) des données, autour d'une valeur centrale, et de trouver un indicateur de cette dispersion.

L'étendue (range) :

Définition 2.1.4 *L'étendue est la différence entre la plus grande et la plus petite valeur observée, elle est notée E :*

$$E = x_{\max} - x_{\min}. \tag{2.17}$$

Elle donne une première idée sur la dispersion des observations.

Les quantiles :

Les quantiles sont la généralisation de la notion de la médiane, qui représente un cas particulier.

Définition 2.1.5 *Un quantile x_α d'ordre α est la valeur de la variable où, α % des observations ont des valeurs qui lui soient inférieure ou égale, et $(1-\alpha)\%$ des observations lui soient supérieure. Il existe trois types de quantiles : les quartiles, les déciles et les centiles.*

1. *Les quartiles* (Q_1, Q_2, Q_3) *:* Ce sont les 3 valeurs qui partagent la série statistique en quatre sous-ensembles d'effectifs égaux. On les notes Q.
2. *Les déciles* $(D_1, ..., D_9)$ *:* Ce sont les 9 valeurs qui partagent la série statistique en dix intervalles d'effectifs égaux. On les notes D.
3. *Les centiles* $(C_1, \ldots, C_{99})$ *:* Ce sont les 99 valeurs qui partagent la série statistique en 100 intervalles d'effectifs égaux, on les notes C.

Les intervalles interquantiles :

Définition 2.1.6 *Les intervalles interquantiles sont l'écart entre le dernier et le premier quantile calculé.*

- *Intervalle interquartile :* $IQ = (Q_3 - Q_1) = x_{\frac{3}{4}} - x_{\frac{1}{4}}$. Contenant 50% d'observations.
- *Intervalle interdécile :* $ID = (D_9 - D_1) = x_{\frac{9}{10}} - x_{\frac{1}{10}}$. Contenant 80% d'observations.
- *Intervalle intercentile :* $IC = (C_{99} - C_1) = x_{\frac{99}{100}} - x_{\frac{1}{100}}$. Contenant 98% d'observations.

Remarque 2.1.3 *L'intérêt des quantiles est l'évaluation des différents intervalles interquantiles.*

Dans la comparaison de la dispersion de deux distributions, plus l'intervalle est important, plus la dispersion est forte.

On calcule les quantiles de même manière du calcul de la médiane.

L'écart absolu moyen :

Définition 2.1.7 *L'écart absolu moyen est la moyenne des distances entre les valeurs observées et leur moyenne en valeur absolue.*

$$e_{moy} = \frac{1}{n}\sum_{i=1}^{p} n_i \left|x_i - \overline{X}\right| \quad ou; \quad e_{moy} = \sum_{i=1}^{p} f_i \left|x_i - \overline{X}\right|. \tag{2.18}$$

Propriété 2.1.3 *1- Si on définit une variable : $Y = aX + b$, où a et b sont des constantes alors : $e_{moy}(Y) = |a| \times e_{moy}(X)$.*

2- $e_{moy}(X) \geq 0$. De plus, $e_{moy}(X) = 0 \Longleftrightarrow x_1 = x_2 = \ldots = x_p$.

Preuve. 1- Pour tout $i = \overrightarrow{1,p}$, on a $y_i = ax_i + b$, et on a $\overline{Y} = a\overline{X} + b$. Par conséquent, $y_i - \overline{Y} = a(x_i - \overline{X})$ et

$$e_{moy}(Y) = \frac{1}{n}\sum_{i=1}^{p} n_i \left|a(x_i - \overline{X})\right| = \frac{1}{n}\sum_{i=1}^{p} |a|\, n_i \left|(x_i - \overline{X})\right| = \frac{1}{n}|a| \sum_{i=1}^{p} n_i \left|(x_i - \overline{X})\right| = |a|\, e_{moy}(X). \tag{2.19}$$

2- Comme $e_{moy}(X)$ est une somme de valeurs absolues divisée par $n > 0$, l'EAM est un rapport de deux nombres non-négatifs. Il est donc non-négatif.

$$e_{moy}(X) = 0 \Longleftrightarrow \sum_{i=1}^{p} n_i \left|(x_i - \overline{X})\right| = 0 \Longleftrightarrow \left|(x_i - \overline{X})\right| = 0, \; i = \overrightarrow{1,p}, \tag{2.20}$$

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 $x_1 = x_2 = \ldots = x_p$.[9] ■

La variance :

Définition 2.1.8 *La variance est la moyenne des carrés des écarts à la moyenne arithmétique. On la symbolise $V(X)$ ou $Var(X)$ ou σ^2*

$$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{p} f_i(x_i - \overline{X})^2. \tag{2.21}$$

Elle mesure la dispersion des modalités de X autour de leur moyenne.

Propriété 2.1.4 *1- Si on a la variable $Y = aX + b$ où, a, b sont des nombres réels quelconques, alors $V(Y) = a^2V(X)$.*
2- $V(X) \geq 0$. On a l'égalité $V(X) = 0 \Longleftrightarrow x_1 = x_2 = \ldots = x_p$.

Preuve. 1- Pour tout $i = \overrightarrow{1,p}$, on a $y_i = ax_i + b$, et on a $\overline{Y} = a\overline{X} + b$. Par conséquent, $y_i - \overline{Y} = a(x_i - \overline{X})$ et

$$V(Y) = \frac{1}{n}\sum_{i=1}^{p} n_i[a(x_i - \overline{X})]^2 = \frac{1}{n}\sum_{i=1}^{p} a^2 n_i(x_i - \overline{X})^2 = \frac{1}{n}a^2\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = a^2V(X). \tag{2.22}$$

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

$$V(X) = 0 \Longleftrightarrow \sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = 0 \Longleftrightarrow (x_i - \overline{X})^2 = 0, i = \overrightarrow{1,p}, \tag{2.23}$$

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 $x_1 = x_2 = \ldots = x_p$.[9] ■

Théorème 2.1.1 (Formule de KÖnig-Huygens)

La variance peut aussi s'écrire

$$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2. \tag{2.24}$$

Preuve. .[8]

$$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i^2 - 2x_i\overline{X} + \overline{X}^2) \tag{2.25}$$
$$= \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - 2 \times \frac{1}{n}\sum_{i=1}^{p} n_i x_i \overline{X} + \frac{1}{n}\sum_{i=1}^{p} n_i \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - 2\overline{X} \times \frac{1}{n}\sum_{i=1}^{p} n_i x_i + \overline{X}^2$$
$$= \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - 2\overline{X}\overline{X} + \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2.$$

■

L'écart-type :

Définition 2.1.9 *L'écart-type est la racine carrée de la variance. On le note* σ_x

$$\sigma_x = \sqrt{V(X)}. \tag{2.26}$$

Remarque 2.1.4 *L'écart-type représente quelque chose très précise pour notre série statistique, il sert à quantifier, mesurer la dispersion d'une série par rapport à sa moyenne. Si on compare deux ou plusieurs distributions d'unités différentes, il est impossible d'utiliser l'écart-type comme indicateur de dispersion, on utilise le coefficient de variation. Plus l'écart-type est petit, plus les données sont concentrées autour de la moyenne. L'écart-type tire toutes les propriétés de la variance. Son unité est même que la variable.*

Coefficient de variation :

Définition 2.1.10 *Le coefficient de variation est le rapport entre l'écart-type et la moyenne exprimée sous forme d'un pourcentage. On le note CV*

$$CV(X) = \frac{\sigma_x}{|\overline{X}|}. \tag{2.27}$$

C'est un nombre sans dimension qui permet de comparer les dispersions de toutes les

distributions différentes, il exprime l'importance de la variabilité par rapport à la valeur centrale.

Remarque 2.1.5 *Le coefficient de variation est l'écart-type de la variable* $\frac{X}{\overline{X}}$ *(il mesure la dispersion de la même façon que l'écart-type).*

Moments :

Définition 2.1.11 *On appelle moment simple d'ordre* $r \in \mathbb{N}$ *de la variable* X*, noté* m_r *la quantité :*

$$m_r = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^r \text{ ou, } m_r = \sum_{i=1}^{p} f_i x_i^r. \tag{2.28}$$

- *Pour* $r = 0 : m_0 = 1$.

- *Pour* $r = 1 : m_1 = \overline{X}$*. C'est la moyenne arithmétique.*

- *Pour* $r = 2 : m_2 = \overline{X}^2 = Q^2$*. C'est le carré de la moyenne quadratique* [1].

- On appelle moment centré d'ordre $r \in \mathbb{N}$ par rapport à une constante a :

$$\mu_a^r = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - a)^r = \sum_{i=1}^{p} f_i (x_i - a)^r. \tag{2.29}$$

– Pour $a = \overline{X}$ et $r = 1 : \mu_{\overline{X}}^1 = 0$. (Ce qui revient à la deuxième propriété de la moyenne arithmétique).

– Pour $a = \overline{X}$ et $r = 2 : \mu_{\overline{X}}^2 = \sigma_x^2 = V(X)$.

- Il existe des relations entre les moments centrés et les moments simples, qui permettent de calculer les premiers à partir des seconds.

$$\begin{cases} \mu_{\overline{X}}^1 = 0. \\ \mu_{\overline{X}}^2 = m_2 - (m_1)^2. \end{cases} \quad \begin{cases} \mu_{\overline{X}}^3 = m_3 - 3m_1m_2 + 2(m_1)^3. \\ \mu_{\overline{X}}^4 = m_4 - 4m_1m_3 + 6(m_1)^2m_2 - 3(m_1)^4. \end{cases}$$

Exemple 2.1.2 *D'après les données de l'exemple 2.1.1, on détermine les paramètres de dispersion suivants :*

[1] $Q = (\frac{1}{n}\sum_{i=1}^{p} n_i x_i^2)^{1/2}$ou, $Q = (\sum_{i=1}^{p} f_i x_i^2)^{1/2}$. La moyenne quadratique.

– *L'étendue :*

$$E = x_{\max} - x_{\min} = x_7 - x_1 = 19 - 0 = 19$$

$$E = 19.$$

– *Les quantiles :* on a $n = 72 \implies \dfrac{n}{4} = 18$.

- A partir de la colonne des N_i, on cherche la valeur $N_i = 18$, tel que la classe correspond à cette valeur est $[8; 10[$

- La valeur exacte de Q_1, on utilise cette relation :

$$\begin{aligned} Q_1 &= b_i + (b_{i+1} - b_i) \times \frac{\frac{n}{4} - N_i}{N_{i+1} - N_i} \\ &= 8 + (10 - 8) \times \frac{18 - 17}{27 - 17} \\ Q_1 &= 8.2. \end{aligned}$$

- $Q_2 = M_e = 11.2$.

- Q_3 : on a $3 \times \dfrac{n}{4} = 54$. La classe correspond à $N_i = 54$ est $[12; 16[$

- La valeur exacte, d'après :

$$\begin{aligned} Q_3 &= b_i + (b_{i+1} - b_i) \times \frac{3 \times \frac{n}{4} - N_i}{N_{i+1} - N_i} \\ &= 12 + (16 - 12) \times \frac{54 - 42}{63 - 42} \\ Q_3 &= 14.28. \end{aligned}$$

$$M_e - Q_1 = 11.2 - 8.2 = 3,$$

$$Q_3 - M_e = 14.28 - 11.2 = 3.08.$$

- On remarque que la distance entre Q_1 et M_e, et la distance entre Q_3 et M_e, est prèsque la même distance qui signifie que la distribution est prèsque symétrique.

– *Intervalle interquantile :*

$$IQ = (Q_3 - Q_1) = 14.28 - 8.2 = 6.08.$$

- Même principe pour les déciles et les centiles.

– *L'écart absolu moyen :*

$$\begin{aligned} e_{moy} &= \frac{1}{n}\sum_{i=1}^{p} n_i \left|c_i - \overline{X}\right| = \frac{1}{72} \times [4 \times |2 - 11.1| + 13 \times |6 - 11.1| \\ &\quad + 10 \times |9 - 11.1| + 15 \times |11 - 11.1| + 21 \times |14 - 11.1| + 9 \times |18 - 11.1| \\ e_{moy} &= 3.44. \end{aligned}$$

– *La variance :*

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{p} n_i (c_i - \overline{X})^2 = \frac{1}{72} \times [4 \times (2 - 11.1)^2 + 13 \times (6 - 11.1)^2 \\ &\quad + 10 \times (9 - 11.1)^2 + 15 \times (11 - 11.1)^2 + 21 \times (14 - 11.1)^2 + 9 \times (18 - 11.1)^2 \\ V(X) &= 18.31. \end{aligned}$$

– *L'écart-type :*

$$\sigma_x = \sqrt{V(X)} = \sqrt{18.31} = 4.27.$$

– *Le coefficient de variation :*

$$CV(X) = \frac{\sigma_x}{\left|\overline{X}\right|} = \frac{4.27}{11.1} = 0.38 = 38\%.$$

- On observe que l'écart-type, la variance et le coefficient de variation sont grands, qui signifie que la dispersion de la distribution est forte. On a donc ce graphe qui explique ces

résultats :

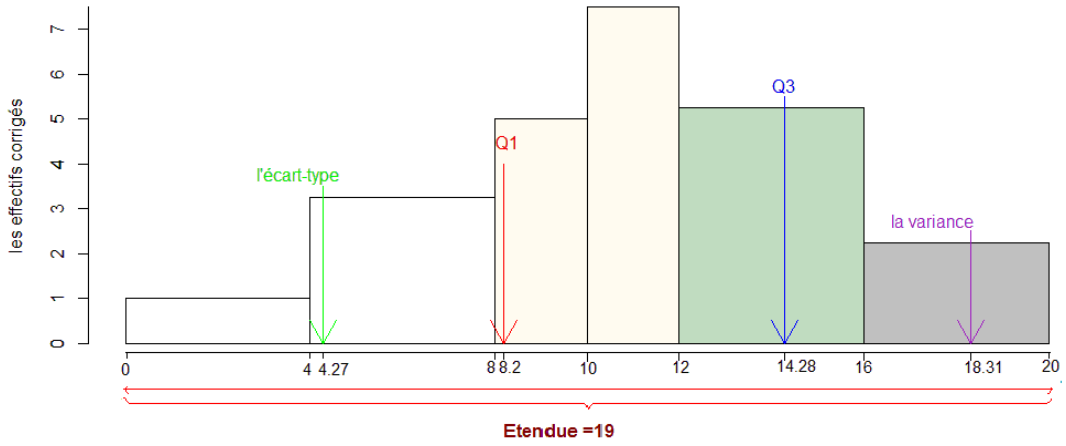

FIG. 2.2 – Histogramme des paramètres de dispersion

2.1.3 Paramètres de forme

Les paramètres de forme permettent de décrire la forme de la distribution statistique par : la symétrique et l'aplatissement, on les définit que pour les variables quantitatives.

Coefficient d'asymétrie :

Définition 2.1.12 *Le coefficient d'asymétrie mesure et compare l'écart de la distribution par rapport à la symétrie.*

Le principe de l'asymétrie est d'évaluer, si la distribution est plus étalée à gauche ou à droite par rapport à une valeur centrale.

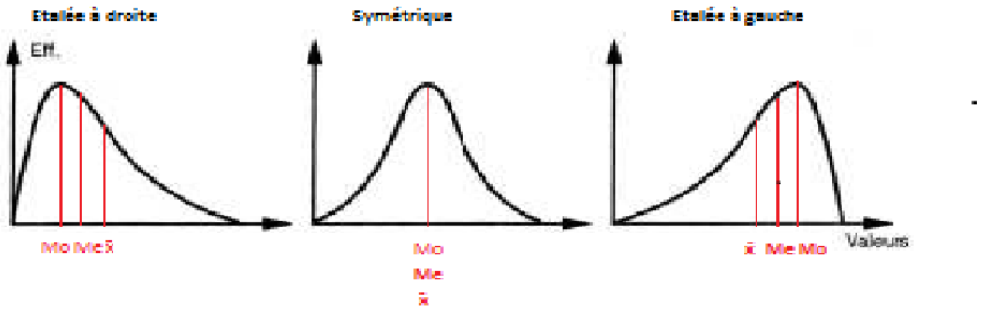

FIG. 2.3 – L'asymétrie d'une distribution

- $\begin{cases} \text{Si } \overline{X} > M_e > M_o \implies \text{la distribution est étalée à droite.} \\ \text{Si } \overline{X} < M_e < M_o \implies \text{la distribution est étalée à gauche.} \\ \text{Si } \overline{X} = M_e = M_o \implies \text{la distribution est symétrique.} \end{cases}$

Il existe plusieurs coefficients d'asymétrie, qui permettent de trouver leur principal intérêt dans le cadre de comparaison de distribution, les principaux sont les suivants :

Le coefficient de Pearson :

Définition 2.1.13 *Le coefficient de Pearson est basé sur une comparaison de la moyenne et du mode. Il s'écrit :*

$$P = \frac{\overline{X} - M_o}{\sigma_x}. \tag{2.30}$$

- $\begin{cases} \text{Si } P > 0 \implies \text{la distribution est étalée à droite.} \\ \text{Si } P < 0 \implies \text{la distribution est étalée à gauche.} \\ \text{Si } P = 0 \implies \text{la distribution est symétrique.} \end{cases}$

Le coefficient de Yule :

Définition 2.1.14 *Le coefficient de Yule est basé sur les positions des trois quartiles. Il*

s'écrit :

$$Y = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1}. \tag{2.31}$$

- $\begin{cases} \text{Si } Q_3 - M_e > M_e - Q_1 \implies \text{la distribution est étalée à droite.} \\ \text{Si } Q_3 - M_e < M_e - Q_1 \implies \text{la distribution est étalée à gauche.} \\ \text{Si } Q_3 - M_e = M_e - Q_1 \implies \text{la distribution est symétrique.} \end{cases}$

Le coefficient de Fisher :

Définition 2.1.15 *Le coefficient de Fisher est basé sur les moments centrés. Il s'écrit :*

$$F = \frac{\mu^3_{\overline{X}}}{(\mu^2_{\overline{X}})^{3/2}} = \frac{\mu^3_{\overline{X}}}{\sigma^3}. \tag{2.32}$$

- $\begin{cases} \text{Si } F > 0 \implies \text{la distribution est étalée à droite.} \\ \text{Si } F < 0 \implies \text{la distribution est étalée à gauche.} \\ \text{Si } F = 0 \implies \text{la distribution est symétrique.} \end{cases}$

Coefficient d'aplatissement (Kurtosis) :

Définition 2.1.16 *Le coefficient d'aplatissement mesure le degré d'aplatissement de la distribution de X. Il concerne la concentration des données observées autour du mode, et la comparer par rapport à la distribution normale.*

FIG. 2.4 – L'aplatissement d'une distribution

- On peut le mesurer avec deux indicateurs :

Le coefficient d'aplatissement de Pearson :

$$\beta_2 = \frac{\mu_X^4}{(\mu_X^2)^2} = \frac{\mu_X^4}{(\mu_X^2)^2}. \tag{2.33}$$

– $\begin{cases} \text{Si } \beta_2 > 3 \implies \text{la distribution est pointue.} \\ \text{Si } \beta_2 < 3 \implies \text{la distribution est aplatie.} \\ \text{Si } \beta_2 = 3 \implies \text{la distribution est normale.} \end{cases}$

Le coefficient d'aplatissement de Fisher :

$$F_2 = \frac{\mu_X^4}{(\mu_X^2)^2} - 3. \tag{2.34}$$

Le coefficient de Pearson prend la valeur $\beta_2 = 3$, quand la distribution normale, donc pour comparer l'aplatissement d'une distribution statistique par l'aplatissement d'une distribution normale, on utilise le coefficient de Fisher : $F_2 = \beta_2 - 3$, tel que :

– $\begin{cases} \text{Si } F_2 = 0 \implies \text{a distribution est normale.} \\ \text{Si } F_2 < 0 \implies \text{la distribution est aplatie.} \\ \text{Si } F_2 > 0 \implies \text{la distribution est pointue.} \end{cases}$

Exemple 2.1.3 *(Pour une variable quantitative discrète)*

Pour une classe de 30 *élèves, on connait le nombre de frères et sœurs de chaque élève. D'après le calcul des effectifs, les ECC, les ECD, les fréquences, les FCC, les FCD, on présente ces données sous forme de tableau suivant :*

Nombre de frères et de sœurs $\mathbf{x}_i$	0	1	2	3	4	5
Effectif $\mathbf{n}_i$	5	10	8	4	1	2
ECC $\mathbf{N}_i$	5	15	23	27	28	30
ECD	30	25	15	7	3	2
Fréquence $\mathbf{f}_i$ (valeur approchée)	0.17	0.33	0.27	0.13	0.03	0.07
(**FCC**) (**valeur approchée**) **$\mathbf{F}_i$**	0.17	0.50	0.77	0.90	0.93	1
(**FCD**) (**valeur approchée**)	1	0.83	0.50	0.23	0.1	0.07

TAB. 2.2 – Tableau de nombre de frères et soeurs d'une classe

1- Déterminer les paramètres caractéristiques.

2- Tracer le diagramme en bâtons.

3- Tracer le diagramme cumulatif.

Solution 2.1.1 ***Les paramètres caractéristiques :***

– **Les paramètres de tendance centrale :**

– *Le mode :*

Le mode est $M_o = x_2 = 1$. Qui signifie que la majorité des élèves ont 1 frère et sœur.

– *La médiane :* comme n est pair ($n = 30$), alors : $\frac{n}{2} = \frac{30}{2} = 15$. Donc :

$$M_e = \frac{1}{2} \times (x_{15} + x_{16}) = \frac{1}{2} \times (1 + 2) = 1.5.$$

$M_e = 1.5$. Cela signifie que la médiane n'est pas nécessairement une des valeurs de X.

– *La moyenne arithmétique :* on la calcule à partir des effectifs :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{6} n_i x_i = \frac{1}{30} \times [(5 \times 0) + (10 \times 1) + (8 \times 2) + (4 \times 3) + (1 \times 4) + (2 \times 5)] = 1.73.$$

$\overline{X} = 1.73$.

- Avec les fréquences :

$$\overline{X} = \sum_{i=1}^{6} f_i x_i = [(0.17 \times 0) + (0.3 \times 1) + (0.3 \times 2) + (0.13 \times 3) + (0.03 \times 4) + (0.07 \times 5)] \simeq 1.73.$$

- On remarque

$$\overline{X} > M_e > M_o.$$

Qui donne une idée sur la position de la distribution, la distribution est étalée à droite.

– **Les paramètres de dispersion :**

– *L'étendue :* l'étendue de cet exemple est :

$$E = x_6 - x_1 = 5 - 0 = 5.\ E = 5.$$

– *Les quantiles :* le principe de calcul des quantiles est le même que la médiane.

– *Les quartiles :*

Le 1$^{\text{er}}$ quartile : on a $n = 30$, $\dfrac{30}{4} = 7.5$.

Dans la colonne des effectifs cumulés croissants, on cherche la modalité correspond à $ECC = 7.5$.

- $Q_1 = x_{\frac{1}{4}} = x_2 = 1$. Cela signifie que 25% des élèves ont 1 frère et sœur ou plus.
- $Q_2 = M_e = 1.5$.
- Le 3$^{\text{ième}}$ quartile : $3 \times \dfrac{n}{4} = 3 \times \dfrac{30}{4} = 22.5$.
- $Q_3 = x_{\frac{3}{4}} = x_3 = 2$. (75% des élèves ont 2 frères et sœurs ou plus).

– *L'intervalle interquartile :*

$$Q_3 - Q_1 = 2 - 1 = 1.$$

– *Les déciles :*

- Le 1$^{\text{er}}$ décile : on a $\dfrac{n}{10} = \dfrac{30}{10} = 3.$

$$D_1 = x_{\frac{1}{10}} = x_1 = 0.$$

- Le 5$^{\text{ième}}$ décile : $D_5 = M_e = 1.5.$
- Le 9$^{\text{ième}}$ décile : on a $9 \times \dfrac{n}{10} = 9 \times \dfrac{30}{10} = 27.$

$$D_9 = x_{\frac{9}{10}} = x_4 = 3.$$

– *L'intervalle interdécile :*

$$D_9 - D_1 = 3 - 0 = 3.$$

– *Les centiles :*

- Le 1$^{\text{er}}$ centile : on a $\dfrac{n}{100} = \dfrac{30}{100} = 0.3.$

$$C_1 = x_{\frac{1}{100}} = x_1 = 0.$$

- Le 50$^{\text{ième}}$ centile : $C_{50} = M_e = 1.5.$
- Le 99$^{\text{ième}}$ centile : on a $99 \times \dfrac{n}{100} = 99 \times \dfrac{30}{100} = 29.7.$

$$C_{99} = x_{\frac{99}{100}} = x_6 = 5.$$

– *L'intervalle intercentile :*

$$C_{99} - C_1 = 5 - 0 = 5.$$

– *L'écart absolu moyen :*

$$\begin{aligned} e_{moy} &= \frac{1}{n}\sum_{i=1}^{6} n_i \left|x_i - \overline{X}\right| \\ &= \frac{1}{30} \times [5 \times |0 - 1.73| + 10 \times |1 - 1.73| + 8 \times |2 - 1.73| \\ &+ 4 \times |3 - 1.73| + 1 \times |4 - 1.73| + 2 \times |5 - 1.73|] \\ e_{moy} &= 1.066. \end{aligned}$$

– *La variance :*

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{6} n_i (x_i - \overline{X})^2 \\ &= \frac{1}{30} \times [5 \times (0 - 1.73)^2 + 10 \times (1 - 1.73)^2 + 8 \times (2 - 1.73)^2 \\ &+ 4 \times (3 - 1.73)^2 + 1 \times (4 - 1.73)^2 + 2 \times (5 - 1.73)^2] \\ V(X) &= 1.8. \end{aligned}$$

- On peut aussi la calculer à partir de

$$\begin{aligned} V(X) &= \frac{1}{n}\sum_{i=1}^{6} n_i x_i^2 - \overline{X}^2 = \frac{1}{30} \times [(5 \times 0^2) + (10 \times 1)^2 + (8 \times 2)^2 \\ &+ (4 \times 3)^2 + (1 \times 4)^2 + (2 \times 5)^2] - (1.73)^2 \\ V(X) &= 1.8. \\ &= \sum_{i=1}^{6} f_i x_i^2 - \overline{X}^2 = [(0.17 \times 0^2) + (0.3 \times 1^2) + (0.3 \times 2^2) \\ &+ (0.13 \times 3^2) + (0.03 \times 4^2) + (0.07 \times 5^2)] - (1.73)^2 \\ V(X) &= 1.8. \end{aligned}$$

– *L'écart-type :*

$$\sigma_x = \sqrt{V(X)} = \sqrt{1.8} = 1.34.$$

– *Coefficient de variation :*

$$C_V = \frac{\sigma_x}{\overline{X}} = \frac{1.34}{1.73} = 0.77 = 77\%.$$

- Le coefficient de variation est grand, qui signifie que la dispersion de la distribution est plus forte.

– **Les paramètres de forme :**

– *Coefficient d'asymétrie de Pearson :*

$$P = \frac{\overline{X} - M_o}{\sigma_x} = \frac{1.73 - 1}{1.34} = 0.54.$$

– *Coefficient d'asymétrie de Yule :*

$$Y = \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1} = \frac{(2+1) - 2 \times 1.5}{1} = 0.$$

– *Coefficient d'asymétrie de Fisher :*

$$\begin{aligned}
F &= \frac{\mu^3_{\overline{X}}}{(\sigma_x)^3}, \\
\mu^3_{\overline{X}} &= \frac{1}{n}\sum_{i=1}^{6} n_i(x_i - \overline{X})^3 \\
&= \frac{1}{30} \times [5 \times (0 - 1.73)^3 + 10 \times (1 - 1.73)^3 + 8 \times (2 - 1.73)^3 \\
&\quad + 4 \times (3 - 1.73)^3 + 1 \times (4 - 1.73)^3 + 2 \times (5 - 1.73)^3] \\
\mu^3_{\overline{X}} &= 2.0069. \\
F &= \frac{2.0069}{(1.34)^3} = 0.834.
\end{aligned}$$

- On remarque que tous les coefficients d'asymétrie sont supérieur ou égal à 0, et comme $\overline{X} > M_e > M_o$, on conclue que la distribution est étalée à droite.

– *Coefficient d'aplatissement de Pearson :*

$$\beta_2 = \frac{\mu_{\overline{X}}^4}{(\mu_{\overline{X}}^2)^2} = \frac{\mu_{\overline{X}}^4}{V(X)^2}.$$

$$\begin{aligned}
\mu_{\overline{X}}^4 &= \frac{1}{n}\sum_{i=1}^{6} n_i(x_i - \overline{X})^4 \\
&= \frac{1}{30} \times [5 \times (0 - 1.73)^4 + 10 \times (1 - 1.73)^4 + 8 \times (2 - 1.73)^4 \\
&+ 4 \times (3 - 1.73)^4 + 1 \times (4 - 1.73)^4 + 2 \times (5 - 1.73)^4] \\
&= 10.44. \\
\beta_2 &= \frac{10.44}{(1.8)^2} = 3.22.
\end{aligned}$$

– *Coefficient d'aplatissement de Fisher :*

$$F_2 = \beta_2 - 3 = 3.22 - 3 = 0.22.$$

- On remarque que $F_2 > 0$, ce qui signifie que la distribution est pointue.

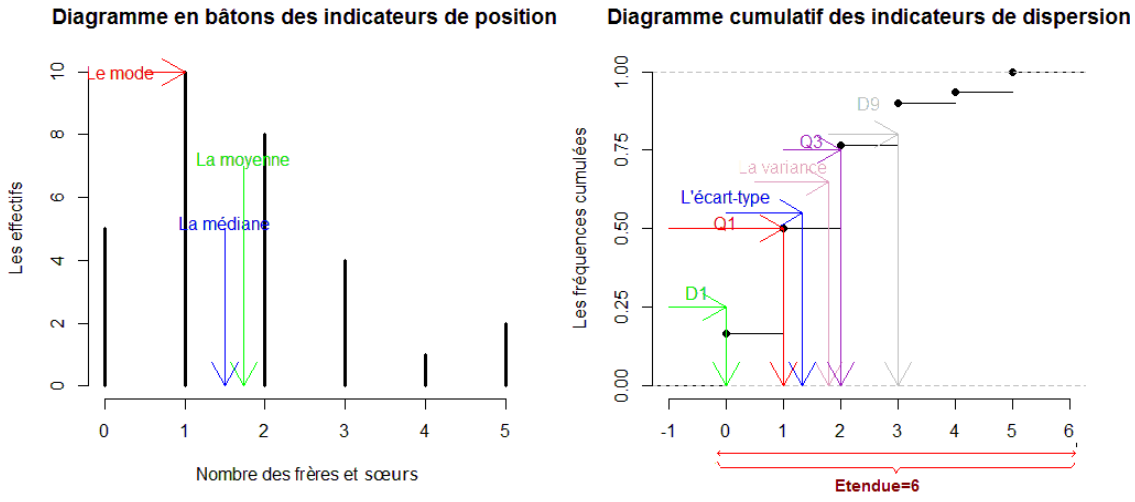

FIG. 2.5 – Les diagrammes des paramètres caractéristiques d'une variable quantitative discrète

Dans le diagramme en bâtons, on ajoute des flèches pour illustrer les paramètres de position, dont le plus haut bâton est le mode qui prend la valeur $x_2 = 1$, qui signifie que la majorité des élèves ont 1 frère et sœur, et la valeur $x_i = 1.5$ signifie la médiane, où 50% des élèves ont un nombre de frères et sœurs entre 1 et 2, et l'autre 50% des élèves ont plus de 2 frères et sœurs. La valeur centrale de cet exemple est $\overline{X} = 1.73$.

Dans la courbe cumulative, on a quantifié les indicateurs de dispersion, qu'ils mesurent l'ordre de grandeur de la distribution de nombre de frères et sœurs autour de la moyenne. On remarque que Q_1 est la valeur, qui découpe une aire représentant 25% de l'aire totale, qui signifie que 25% des élèves ont 1 frère et sœur ou plus, et 75% des élèves ont plus de 1 frère et sœur, et Q_3 découpe une aire représentant 75% de l'aire totale, qui signifie que 75% des élèves ont 2 frères et sœurs ou plus, et 25% des élèves ont plus de 2 frères et sœurs.

On observe aussi, que les valeurs observées sont concentrées autour de la moyenne, cela signifie que la dispersion de la distribution est petite.

Conclusion

Dans ce mémoire, on a vu un bref résumé sur la statistique descriptive univariée, qui sert 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 la technique numérique (qui donne des interprétations de résultats obtenues).

Tous les informations et les résultats obtenues concernant les deux techniques précédentes, sont utilisées pour l'étude de la statistique descriptive, mais contrairement à la statistique descriptive, on parle sur la statistique inférentielle qui ne se contente pas de décrire des observations, mais extrapole les constatations faites à un ensemble plus vaste, permet de tester des hypothèses sur cet ensemble et de prendre des décisions le concernant. La statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données.

A la fin, la statistique descriptive univariée est une méthode principale dans l'étude statistique des phénomènes.

Bibliographie

[1] Baccini, A. Statistique Descriptive Élémentaire.

[2] Chekroun. A. (2017 – 2018). Statistiques descriptives et exercices.

[3] Goga. C et Labruère. C. (2009). Intruduction au logiciel R. Ecole Doctorale Dijon.

[4] Hammdani, H. (1988). Statistique descriptive et expression graphique.

[5] Hubler, J. (2007). Statistique descriptive appliquée à la gestion et à l'économie. Editions Bréal.

[6] IMMEDIATO. H. LICENCE Scientifique. Cours. Statistiques.

[7] Khaldi, K. (2001). Méthodes statistiques et probabilités.

[8] Tille. Y. (2010). Résumé du Cours de Statistique Descriptive.

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

Annexe B : Abréviations et Notations

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

ECC	Effectifs cumulés croissants.
ECD	Effectifs cumulés décroissants.
FCC	Fréquences cumulées croissantes.
FCD	Fréquences cumulées décroissantes.
$\sum$	Le symbole sigma est une notation de la somme des valeurs.
EAM	L'écart absolu moyen.
$\mathbb{N}$	L'ensemble des nombres entiers naturels.
$card(\Omega)$	Le cardinal : nombre d'éléments de l'ensemble Ω.
$\sum_{i=1}^{p}$	La somme pour i variant de 1 à p.