République Algériennne 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

Départementde Mathématiques

Mémoire présenté pour obtenir le diplôme de

Master en "**Mathematiques Appliquées**"

Option : **Statistique**

Par

FELATA Sana

Titre :

La statistique descriptive univariée

Membres du Comité d'Examen :

Pr.	**CHERFAOUI Mouloud**	UMKB	Président
Pr.	**SAYAH Abdallah**	UMKB	Encadreur
Dr.	**BENELMIR Imane**	UMKB	Examinateur

Juin 2023

Dédicace

Je dédie ce humble travail à :

Mes chers parents

Tous les membres de ma famille :

Ma chère soeur "Racha", et son Mari "Soulbi Salah"

Mes troix frères :

"Amin", "Amar", "Bachir"

Ma chère amie :

Rebiha

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

et en particulier Abdelhakim Necir, Djabrane Yahia.

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 : **SAYAH ABDALLAH**

Et lui souhaite le meilleur dans sa carrière professionnelle.

Je tiens à remercier : **CHERFAOUI Mouloud**, **BENELMIR Imane**,

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

1.1 Définitions 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 caractére, 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).

1.2 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. 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 : 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, 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), 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,...

1.3 Types de la variable statistique

II existe deux types de caractères : les caractères qualitatifs et les caractères quantitatifs.

1.3.1 Caractères qualitatifs

Variables nominales et variables ordinales

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées modalités. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite ordinale. Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite nominale.

1.3.2 Caractère quantitatif

Cas d'une variable quantitative discrète

En général, on appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs entières (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible (sauf exception, moins d'une vingtaine). Citons, par exemple, le nombre d'enfants dans une population de familles, le nombre d'années d'études après le bac dans une population

d'étudiants...

Variable statistique 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 $\mathbb{R}$ $([b_i; b_{i+1}[)$.

1.4 Effectifs, fréquences, fréquences cumulée

1.5 Effectif partiel

Définition 1.5.1 *Pour chaque valeur x_i, on pose par définition*

$$n_i = \text{card}\left\{\omega \in \Omega : X(\omega) = x_i\right\} \qquad i = 1...k \tag{1.1}$$

n_i : le nombre d'individus qui ont le même x_i, ça s'appelle effectif partiel de x_i.

1.5.1 Effectif total

Définition 1.5.2 *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*

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

1.5.2 Fréquence partielle (fréquence relative)

Définition 1.5.3 *Pour chaque valeur x_i, on pose par définition f_i*

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

f_i s'appelle la fréquence partielle de x_i. La fréquence d'une valeur est le rapport de l'effectif de cette valeur par l'effectif total.

Remarque 1.5.1 ***1)*** *La somme des fréquences relatives est égale à 1 car la valeur de la fréquence est toujours entre 0 et 1.*

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

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

1.5.3 Effectif et fréquence cumulé croissant

Définition 1.5.4 *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 :*

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

1.5.4 Effectif et fréquence cumulé décroissant

Définition 1.5.5 *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 :*

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

1.6 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.

1.6.1 Série statistique

On appelle série statistique la suite des valeurs prises par une variable X sur les unités d'observation.

Le nombre d'unités d'observation est noté n.

Les valeurs de la variable X sont notées $x_1, ..., x_i, ..., x_n$.

1.6.2 Tableau statistique

On appelle donc tableau statistique un tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X. Ces observations sont

rangées par ordre croissant et non répétées ; nous les noterons$\{x_i; i = 1; ...; r\}$.

Tableau statistique d'une variable qualitative nominale

Exemple 1 :

On s'intéresse à la variable statistique nominale état-civil (Célibataire, Marié, Divorcé, Veuf) notée X et à la série statistique des valeurs prises par X sur 25 personnes qui est représentée par le tableau suivant :

x_i	n_i	f_i
Célibataire	5	0.2
Marié	11	0.44
Divorcé	6	0.24
Veuf	3	0.12
Total	25	1.00

tableau statistique suivant :

x_i	n_i	$N_i = ECC$	f_i	$F_i = FCC$	$N_i = ECD$	$F_i = FCD$
0	5	5	0.05	0.05	100	1
1	19	24	0.19	0.24	95	0.95
2	27	51	0.27	0.51	76	0.76
3	21	72	0.21	0.72	49	0.49
4	15	87	0.15	0.87	28	0.28
5	9	96	0.09	0.96	13	0.13
6	4	100	0.04	1	4	0.04
Tatal	100		1			

Tableau statistique d'une variable qualitative ordinale

Exemple 2 :

On interroge 60 personnes sur leur niveau d'étude, et on s'intéresse à la variable statistique ordinale (Sans niveau, Niveau primaire, moyen, secondaire, universitaire) notée Y et à la série statistique des valeurs prises par Y sur 60 personnes qui est définie par le tableau suivant

x_i	n_i	$N_i = ECC$	f_i	$F_i = FCC$	$N_i = ECD$	$F_i = FCD$
Sans niveau	3	3	0.05	0.05	60	1
Niveau primaire	9	12	0.15	0.2	57	0.95
Niveau moyen	12	24	0.2	0.4	48	0.80
Niveau secondaire	21	45	0.35	0.75	36	0.60
Niveau universitaire	15	60	0.25	1.00	15	0.25
Total	60		1			

Remarque :

Si la variable est ordinale, on peut calculer les effectifs cumulés et les fréquences cumulées.

Caractère quantitatif discret :

Les modalités x_i	**Effectif n_i**	**ECC N_i**	**Fréquence F_i**	**FCC F_i**
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
.....	.....	.....	.....	.....
x_p	n_p	N_p	f_p	F_p
Total	N		1	

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

Tableau statistique d'une variable quantitative discrète

Un quartier est composé de 100 familles, et la variable Z représente le nombre d'enfants par famille. Les valeurs de la variable sont

x_i	0	1	2	3	4	5	6
n_i	5	19	27	21	15	9	4

Pour les variables quantitatives discrètes, on peut calculer les effectifs, les effectifs cumulés croissants et décroissants, les fréquences, les fréquences cumulées croissantes et décroissantes, donc le tableau statistique est le suivant

x_i	n_i	$N_i = ECC$	f_i	$F_i = FCC$	$N_i = ECD$	$F_i = FCD$
0	5	5	0.05	0.05	100	1
1	19	24	0.19	0.24	95	0.95
2	27	51	0.27	0.51	76	0.76
3	21	72	0.21	0.72	49	0.49
4	15	87	0.15	0.87	28	0.28
5	9	96	0.09	0.96	13	0.13
6	4	100	0.04	1	4	0.04
Tatal	100		1			

Caractère quantitatif continu

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

TAB. 1.2 – Tableau statistique regroupé par classes

On commence par les notations suivantes

Soit la classe de la modalité $b_i = [b_p; b_{p+1}[$, on note, de manière générale :

$\mathbf{b}_p$ la borne inférieure de la classe b_i,

$\mathbf{b}_{p+1}$ la borne supérieure de la classe b_i,

$\mathbf{c}_i = \dfrac{b_p + b_{p+1}}{2}$ le centre de la classe b_i,

$\mathbf{a}_i = \mathbf{b}_{p+1} - \mathbf{b}_p$ l'amplitude de la classe b_i

$\mathbf{d}_i = \dfrac{n_i}{a_i}$ la densité d'effectif (ou effectif unitaire) de la classe b_i,

$\delta_i = \dfrac{f_i}{a_i}$ la densité de fréquence (ou fréquence unitaire) de la classe b_i.

Exemple 4 :

En mesurant la taille en centimètres de 40 étudiants d'une classe on trouve

x_i	159	160	161	162	163	164	165	166	167	168
n_i	3	5	2	4	1	7	8	6	1	3

Pour les variables quantitatives continues, le tableau statistique est basé sur la désignation des classes $b_i = [b_p; b_{p+1}[$, puis on calcule le centre de chaque classe, les effectifs, la densité d'effectif, les effectifs cumulés, les fréquences, la densité de fréquence et les fréquences cumulées, d'où le tableau statistique est le suivant :

Classes	Centre de la classe	n_i	$\mathbf{d}_i$	N_i	f_i	δ_i	F_i
$[158.5, 160.5[$	159.5	8	4	8	0.2	0.1	0.2
$[160.5, 162.5[$	161.5	6	3	14	0.15	0.075	0.35
$[162.5, 164.5[$	163.5	8	4	22	0.2	0.1	0.55
$[164.5, 166.5[$	165.5	14	7	36	0.35	0.175	0.9
$[166.5, 168.5[$	167.5	4	2	40	0.10	0.05	1
Total		40			1		

Remarque :

La question qui se pose est en combien de classes partageons-nous les valeurs ?.

Les réponses sont les suivantes :

1) Soit p le nombre d'observations, alors le nombre de classes d'après la formule de Sturge est $k = 1 + 3.3 \log_{10} p$.

2) D'après la formule de Yule, $k = 2.5\sqrt[4]{p}$.

L'intervalle de classe est obtenue de la façon suivante : longueur de l'intervalle est égale à $\dfrac{p-1}{k}$

Si on prend l'exemple précédent on a $p = 10$. Alors le nombre de classes possibles est

$k = 1 + 3.3 \log_{10} 10 = 4.3 \simeq 5$ d'après la formule de Sturg,

ou $k = 2.5\sqrt[4]{10} = 4.4457 \simeq 5$ d'après la formule de Yule.

La longueur de l'intervalle est égale à $\dfrac{10-1}{4.3} = 2.093 \simeq 2$.

Exemple 5 :

On dispose des résultats d'une enquête concernant les loyers annuels des appartements dans une cité d'une ville.

Montant du loyer (x 10000 DA)	n_i
$[5, 7[$	30
$[7, 9[$	15
$[9, 12[$	36
$[12, 16[$	9
$[16, 20[$	12
$[20, 30[$	18
Total	120

Compléter le tableau statistique (centre de chaque classe, les effectifs, la densité d'effectif, les effectifs cumulés, les fréquences, la densité de fréquence et les fréquences cumulées)

Montant du loyer (x 10000 DA)	Centre de la classe	n_i	d_i	N_i	f_i	δ_i	F_i
$[5, 7[$	6	30	15	30	0.25	0.125	0.25
$[7, 9[$	8	15	07.5	45	0.125	0.0625	0.375
$[9, 12[$	10.5	36	12	81	0.3	0.1	0.675
$[12, 16[$	14	9	2.25	90	0.075	0.01875	0.75
$[16, 20[$	18	12	3	102	0.1	0.025	0.85
$[20, 30[$	25	18	1.8	120	0.15	0.015	1
Total		120			1		

1.7 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.

1.7.1 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 classes (où l'amplitude de ces classes est le largeur des rectangles), et leurs surface proportionnelle à l'effectif ou à la fréquence de la classe. Si les classes ne sont pas d'égale amplitude[1], il faut calculer l'hauteur du rectangle comme suit :

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

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

Remarque 1.7.1 *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

[1] L'amplitude d'une classe est : $\mathbf{a}_i = \mathbf{b}_{i+1} - \mathbf{b}_i$.

partir de la fonction de répartition empirique.[2].

Exemple 1.7.1 *Un quartier est composé de* 50 *ménages, et la variable Z représente le nombre de personnes par ménage. Les valeurs de la variable sont*

1	1	1	1	1	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	6	6	6	8	8

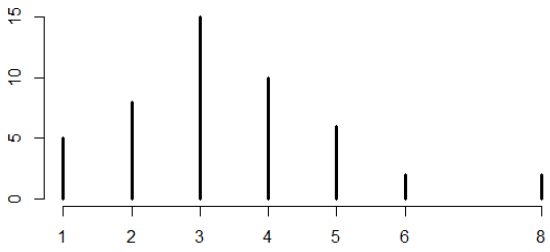
TAB. 1.3 – Tableaue nombre de personnes par ménage

Comme pour les variables qualitatives ordinales, on peut calculer les effectifs, les effectifs cumulés, les fréquences, les fréquences cumulées. à nouveau, on peut ,construire le tableau statistique :

x_j	n_j	N_j	f_j	F_j
1	5	5	0.10	0.10
2	9	14	0.18	0.28
3	15	29	0.30	0.58
4	10	39	0.20	0.78
5	6	45	0.12	0.90
6	3	48	0.06	0.96
8	2	50	0.04	1.00
	50		1	

TAB. 1.4 – le tableau statistique

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



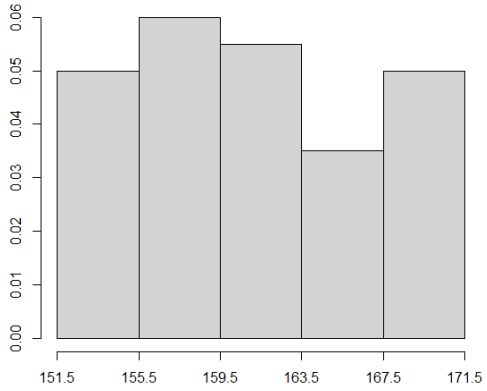
Digramme en bâtonnets des effectifs

Exemple 1.7.2 *(D'une variable quantitative continue)*

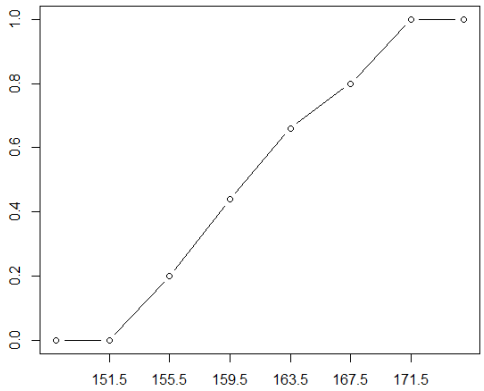
On mesure la taille en centimètres de 50 élèves d'une classe.On obtient le tableau suivant :

Classes	$\mathbf{n}_i$	$\mathbf{c}_i$	$\mathbf{N}_i$	**ECC**	**ECD**	$\mathbf{f}_i$	**FCC**	**FCD**	$\mathbf{n}'_i = \frac{n_i}{a_i}$	$\mathbf{f}' = \frac{\mathbf{f}_i}{a_i}$
$[151.5; 159.5[$	10	153.5	10	10	50	0.20	0.20	1	1.5	0.03
$[155.5; 159.5[$	12	157.5	22	22	40	0.24	0.44	0.8	1.4	0.028
$[159.5; 163.5[$	11	161.5	33	33	28	0.22	0.66	0.56	2.2	0.044
$[163.5; 167.5[$	7	165.5	40	40	17	0.14	0.8	0.34	2.8	0.056
$[167.5; 171.5[$	10	169.5	50	50	10	0.20	1	0.2	0.3	0.006
Total	50					1				

TAB. 1.5 – Tableau de la taille d'élèves



Histogramme des fréquences



Fonction de répartition d'une distribution groupée

1.7.2 Cas d'une variable qualitative

1) Dans le cas d'une variable qualitative, le tableau statistique peut être représenté par trois types de représentation graphique :

le diagramme en barres, le diagramme en bâtons et le diagramme circulaire.

Pour **le diagramme en barres** : on prend en abscisses les modalités de façon arbitraire, et en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs (ou aux fréquences) de façon arbitraire de chaque modalité.

Pour **le diagramme en bâtons** : même méthode que représentation par diagramme en barres.

Pour **le diagramme circulaire :** on partage un disque en secteurs angulaires, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif (ou à la fréquence) de la modalité.

Le degrés d'un secteur circulaire est déterminé pour les effectifs à l'aide de la règle

de trois suivante :

$$N \rightarrow 360^\circ$$

$$n_i \rightarrow b_i \text{ (degrés de la modalité).}$$

Donc pour les effectifs

$$b_i = \frac{n_i}{N} \times 360^\circ.$$

et pour les fréquences

$$b_i = f_i \times 360^\circ$$

Chapitre 2

Statistique descriptive univariée

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 **??**. *Il est noté :*M_0 *ou* x_M. *Et on le calcule numériquement comme suit :*

$$M_0 = x_i + (x_{i+1} - x_i) \times \frac{\left(n'_{i+1} - n'_i\right)}{\left(n'_{i+1} - n'_i\right) + \left(n'_{i+1} - n'_{i+2}\right)}. \tag{2.1}$$

où, x_i *est la borne inferieure de la classe modale,*

x_{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 $[x_i; x_{i+1}[$ est associée au rectangle le plus haut. Le mode se calcule pour tous les types de variables.

Remarque 2.1.1 *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 *La médiane, désignée par* M_e, *est la valeur de la variable qui correspond à 50% des observations.*

La médiane partage donc la série des observations en deux ensembles d'éffectifs 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 : $x_{(\frac{(N+1)}{2})}$.

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

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 associée 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 :

$$\frac{M_e - x_i}{x_{i+1} - x_i} = \frac{\frac{N}{2} - N_i}{N_{i+1} - N_i}. \tag{2.4}$$

$$M_e = x_i + (x_{i+1} - x_i) \times \frac{\frac{N}{2} - N_i}{N_{i+1} - N_i},$$

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

x_{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) superieur 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.

– **Pour les variables continues :**

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

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

Propriété 2.1.1 *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.

On montre que : $\overline{Z} = a\overline{X} + b$.[7]

Preuve. *On a :*

$$\begin{aligned} \overline{Z} = \frac{\sum_{i=1}^{p} n_i z_i}{N} &= \frac{1}{N}\sum_{i=1}^{p} n_i \left(ax_i + b\right) \\ &= \frac{1}{N}\sum_{i=1}^{p} a n_i x_i + \frac{1}{N}\sum_{i=1}^{p} n_i b \end{aligned} \tag{2.9}$$

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 :

$$\frac{a}{N}\sum_{i=1}^{p} n_i x_i + \frac{b}{N}\sum_{i=1}^{p} n_i \tag{2.10}$$

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

$$\frac{a}{N}\sum_{i=1}^{p} n_i x_i + b \times \frac{N}{N} * \tag{2.11}$$

$$\overline{Z} = a\overline{X} + b. \tag{2.12}$$

■

Propriété 2.1.2 *2-* La moyenne des écarts à la moyenne est nulle :

On montre que

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

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 :

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

et comme $\sum_{i=1}^{p} n_i x_i = N\overline{X}$

on a alores :

$$\sum_{i=1}^{p} n_i \left(x_i - \overline{X}\right) = N\overline{X} - N\overline{X} = 0. \tag{2.16}$$

et l'égalité est vérifiée.[4] ■

Exemple 2.1.1 *Pour 20 élèves des classes différentes, on connait les résultats d'un examan d', on obtient le tableau suivant :*

$\mathbf{X}_i$	$\mathbf{n}_i$	$\mathbf{a}_i$	$\mathbf{c}_i$	$\mathbf{N}_i$	f_i	$\mathbf{n}_i^{'}$
$[0;2[$	2	1	2	2	0.1	1
$[2;5[$	2	3.5	3	4	0.1	0.66
$[5;8[$	3	6.5	3	7	0.15	1
$[8;10[$	4	9	2	11	0.2	2
$[10;14[$	5	12	4	16	0.25	1.25
$[14;16[$	3	15	2	19	0.15	1.5
$[16;20[$	1	18	4	20	0.05	0.25
Total	20				1	

TAB. 2.1 – Tableau statistique

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

1. *Le mode :*

   - On cherche d'abord la classe modale, qui est égale à $[8;10[$ (comme les classes n'ont pas d'amplitudes égaux), qui correspond à l'effectif corrigé le plus grand, $n_i' = 2$.

   - Puis, on précise la valeur de mode est :

$$\begin{aligned} M_o &= x_i + (x_{i+1} - x_i) \times \frac{(n_{i+1}^{'} - n_i^{'})}{(n_{i+1}^{'} - n_i^{'}) + (n_{i+1}^{'} - n_{i+2}^{'})} \\ &= 8 + (10 - 8) \times \frac{(2-1)}{(2-1) + (2 - 1.25)} \\ &= 9.14. \end{aligned}$$

2. *La médiane :*

   - On a $N = 20$ est pair, telque $\frac{N}{2} = 10$, alors la classe médiane est la classe qui contient la $10^{ième}$ valeur et la $11^{ième}$ valeur, d'aprés la colonne N_i

on cherche l'intervalle qui correspond à ces valeurs, on trouve que la classe médiane est $[8; 10[$

- La valeur prise est :

$$\begin{aligned} M_e &= x_i + (x_{i+1} - x_i) \times \frac{\frac{N}{2} - N_i}{N_{i+1} - N_i} \\ &= 8 + (10 - 8) \times \frac{10 - 7}{16 - 7} \\ &= 8.75. \end{aligned}$$

3. *La moyenne :*

$$\begin{aligned} \overline{X} &= \frac{\sum_{i=1}^{p} n_i c_i}{N} \\ &= 1/20 \times [(2 \times 1) + (2 \times 3.5) + (3 \times 6.5) + (4 \times 9) + (5 \times 12) + (3 \times 15) + (1 \times 18)] \\ &= 9.375. \end{aligned}$$

On observe que :

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

2.1.2 Paramètres de dispersion

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

L'étendue (range)

Soit X une variable statistique réelle discrète.

L'étendue E de X est la différence entre la plus grande valeur de X et la plus

petite valeur de X

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

Elle est souvent utilisée dans les contrôles industriels en raison de la simplicité, limité et de la rapidité de son calcul.

Les quantiles

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

Définition 2.1.4 *Un quantile x_α d'ordre α est la valeur de la variable où, α % des observations prennent une valeur qui lui soit inferieure. 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 population (ou l'étendue) en quatre sous-ensembles d'effectifs égaux. On les notes Q.
2. *Les déciles ($D_{1,\dots}, D_9$) :* ce sont les 9 valeurs qui partagent l'étendue en dix intervalles d'effectifs égaux. On les notes D.
3. *Les centiles ($C_1, \dots, C_{99}$) :* ce sont les 99 valeurs qui partagent l'étendue en 100 intervalles d'effectifs égaux, on les notes C.

Les intervalles interquantiles

Définition 2.1.5 – *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.*

L'écart absolu moyen

Définition 2.1.6 *L'écart absolu moyen est la moyenne des distances entre les valeurs observées et leur moyenne en valeur absolue.*

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

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

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

$$= |a|\, e_{moy}(X). \tag{2.20}$$

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. Elle est donc non-négative.

$$e_{moy}(X) = 0 \Longleftrightarrow \sum_{i=1}^{p} n_i \left|(x_i - \overline{X})\right| = 0 \Longleftrightarrow \left|(x_i - \overline{X})\right| = 0, i = \overrightarrow{1,p}, \tag{2.21}$$

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 équivalenteà $x_1 = x_2 = \ldots = x_p$.[9] ■

La variance

Définition 2.1.7 *La variance est la moyenne des carrés des écarts à la moyenne arithmétique. On la symbolise $V(X)$ ou $Var(X)$ ou σ^2*

$$Var(X) = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \sum_{i=1}^{p} f_i(x_i - \overline{X})^2. \tag{2.22}$$

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

Propriété 2.1.3 *1. Si on a la variable $Y = aX + b$ où a, b sont des nombres réels quelconques, alors $Var(Y) = a^2 Var(X)$.*

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

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

$$Var(Y) = \frac{1}{N}\sum_{i=1}^{p} n_i[a(x_i - \overline{X})]^2 = \frac{1}{N}\sum_{i=1}^{p} a^2 n_i(x_i - \overline{X})^2 = \frac{1}{N} \times a^2 \sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = a^2 Var(X). \tag{2.23}$$

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

$$Var(X) = 0 \Longleftrightarrow \sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = 0 \Longleftrightarrow (x_i - \overline{X})^2 = 0, i = \overrightarrow{1,p}, \tag{2.24}$$

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*

$$Var(X) = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2. \tag{2.25}$$

■

Preuve. [8]

$$\begin{aligned} Var(X) &= \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \overline{X})^2 = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 + n_i \overline{X}^2 - 2x_i n_i \overline{X} \\ &= \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2\overline{X} \times \frac{1}{N}\sum_{i=1}^{p} x_i n_i + \overline{X}^2. \\ &= \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2\overline{X}^2 + \overline{X}^2 = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2 \end{aligned} \tag{2.26}$$

■

L'écart-type

Définition 2.1.8 *L'écart-type est la racine carrée de la variance*

$$\sigma_x = \sqrt{Var\,(x)} \tag{2.27}$$

Remarque 2.1.3 *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 2.1.9 *Le coefficient de variation est le rapport entre l'écart-type et la moyenne éxprémé sous forme d'un pourcentage .*

$$CV = \frac{\sigma_x}{\left|\overline{X}\right|}. \tag{2.28}$$

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

$$m_p = \frac{1}{N}\sum_{i=1}^{k} n_i \left(x_i - \alpha\right)^p = \sum_{i=1}^{k} f_i \left(x_i - \alpha\right)^p \tag{2.29}$$

On peut définir deux types de moments en fonction de α *:*

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

$$m_p = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^p = \sum_{i=1}^{k} f_i x_i^p \tag{2.30}$$

-pour $p = 0 \Rightarrow m_0 = 1$,

-pour $p = 1 \Rightarrow m_1 = \sum_{i=1}^{k} f_i x_i = \overline{X}$.

-pour $p = 2 \Rightarrow m_2 = \sum_{i=1}^{k} f_i x_i^2 = \overline{X}^2 = mq$ *c'est la moyenne quadratique tel que :*$mq = \left(\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2\right)^2$.

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

$$\mu_{\overline{X}}^{p} = \frac{1}{N}\sum_{i=1}^{k}\left(x_i - \overline{X}\right)^p = \sum_{i=1}^{k} f_i\left(x_i - \overline{X}\right)^p. \tag{2.31}$$

- *Pour* $p = 0 \Rightarrow \mu_{\overline{X}}^{0} = 1$

-Pour $p = 1 \Rightarrow \mu_{\overline{X}}^{1} = 0$.

-Pour $p = 2 \Rightarrow \mu_{\overline{X}}^{2} = var\,(x)$

Relations entre les moment centrés et les moment non centrés :

$\mu_{\overline{X}}^{2} = m_2 - (m_1)^2$.

$\mu_{\overline{X}}^{3} = m_3 - 3m_1m_2 + 2(m_1)^3$.

$\mu_{\overline{X}}^{4} = m_4 - 4m_1m_3 + 6(m_1)^2 m_2 - 3(m_1)^4$

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

1. *L'étendue :*

$$E = c_k - x_1 = 18 - 1 = 17.$$

2. *Les quantiles :* on a $N = 20 \implies \frac{N}{4} = 5$.

   - A partir de la colonne des N_i, on cherche la valeur $N_i = 5$, telque la classe correspond à cette valeur est $[5; 8[$.

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

$$\begin{aligned} Q_1 &= x_i + (x_{i+1} - x_i) \times \frac{\frac{N}{4} - N_i}{N_{i+1} - N_i} \\ &= 5 + (8-5) \times \frac{5-4}{7-4} \\ &= 5.99. \end{aligned}$$

- $Q_2 = M_e = 9.375$.

- Q_3 : on a $\frac{3N}{4} = 15$. La classe correspond à $N_i = 15$ est $[10; 14[$

- La valeur exacte, d'aprés :

$$\begin{aligned} Q_3 &= x_i + (x_{i+1} - x_i) \times \frac{\frac{3N}{4} - N_i}{N_{i+1} - N_i} \\ &= 10 + (14-10) \times \frac{15-11}{16-11} \\ &= 13.2. \end{aligned}$$

$$M_e - Q_1 = 9.375 - 5.99 = 3.385$$

$$Q_3 - M_e = 13.2 - 9,375 = 3.825.$$

- 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 singnifie que la distribution est prèsque symétrique.

1. Intervalle interquantile

$$IQ = (Q_3 - Q_1) = 13.2 - 5.99 = 7.21.$$

- Même principe pour les déciles et les centiles

2. *L'écart absolu moyen :*

$$e_{moy} = \frac{1}{N}\sum_{i=1}^{p} n_i \left|c_i - \overline{X}\right| = 1/20 \times [2 \times |1 - 9.375| + 2 \times |3.5 - 9.375|$$
$$+ 3 \times |6.5 - 9.375| + 4 \times |9 - 9.375| + 5 \times |12 - 9.375| + 3 \times |15 - 9.375| + 1 \times |18 - 9.$$
$$= 3.862.$$

3. *La variance :*

$$Var(X) = \frac{1}{N}\sum_{i=1}^{p} n_i (c_i - \overline{X})^2 = 1/20 \times [2 \times (1 - 9.375)^2 + 2 \times (3.5 - 9.375)^2$$
$$+ 3 \times (6.5 - 9.375)^2 + 4 \times (9 - 9.375)^2 + 5 \times (12 - 9.375)^2 + 3 \times (15 - 9.375)^2 + 1$$
$$= 21.921.$$

4. *L'écart-type :*

$$\sigma_x = \sqrt{Var(X)} = \sqrt{21.921} = 4.68.$$

5. *Le coefficient de variation :*

$$CV(X) = \frac{\sigma_x}{\left|\overline{X}\right|} = \frac{4.68}{9.375} = 0.49 = 49\%.$$

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

2.1.3 Paramètres de forme

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

Coefficient d'asymétrie

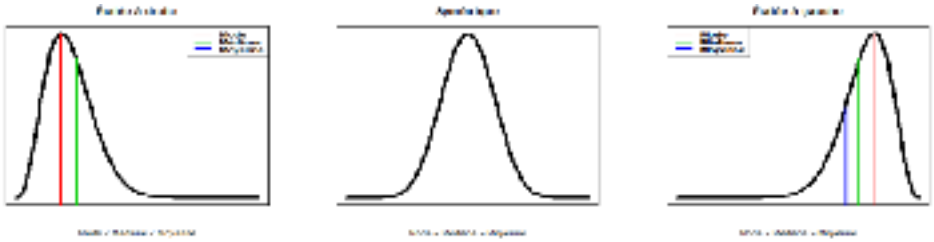


FIG. 2.1 – L'asymétrie d'une distribution

Le coefficient de Pearson

Définition 2.1.11 *Le coefficient de Pearson est basé sur une comparaison de la moyenne et du mode. Il s'écrit :*

$$P = \frac{\overline{X} - M_o}{\sigma_x}. \tag{2.32}$$

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

Le coefficient de Yule

Définition 2.1.12 *Le coefficient de Yule est basé sur les positions des trois quartiles ou, des déciles. Il s'écrit :*

$$Y = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1}. \tag{2.33}$$

efficient de Fisher

Définition 2.1.13 *Le coefficient de Fisher est basé sur les moments centrés. Il s'écrit :*

$$F = \frac{\mu_{\overline{X}}^3}{(\mu_{\overline{X}}^2)^{3/2}} = \frac{\mu_{\overline{X}}^3}{\sigma^3}. \tag{2.34}$$

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

Coefficient d'aplatissement (Kurtosis)

Définition 2.1.14 *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.*

- On peut le mesurer avec deux indicateurs :

– **Le coefficient d'aplatissement de Pearson est :**

$$\beta_2 = \frac{\mu_{\overline{X}}^4}{(\mu_{\overline{X}}^2)^2} \tag{2.35}$$

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

– **Le coefficient d'aplatissement de Fisher est :**

$$F_2 = \frac{\mu_{\overline{X}}^4}{(\mu_{\overline{X}}^2)^2} - 3. \tag{2.36}$$

- 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'aplatis-

sement d'une variable de Gauss, on utilise le coefficient Fisher : $F_2 = \beta_2 - 3$, telque :

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

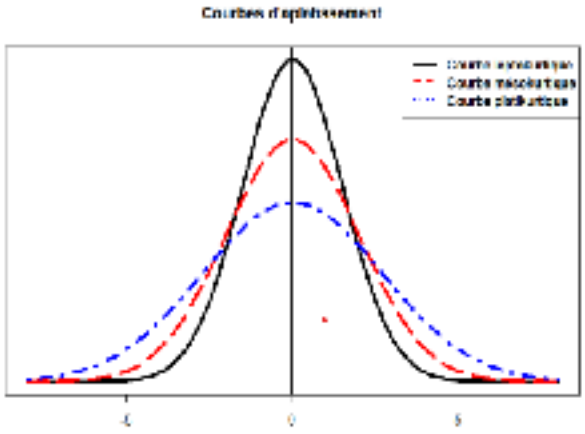


FIG. 2.2 – Aplatissements comparés.

(Pour une variable quantitative discrète) Pour une classe de 30 élèves, on connait le nombre de frères et sœurs de chaque élève. D'aprés le calcule 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 x_i	0	1	2	3	4	5
Effectif n_i	5	10	8	4	1	2
ECC N_i	5	15	23	27	28	30
ECD N_i	30	25	15	7	3	2
Fréquence f_i (valeur approchée)	0.17	0.33	0.27	0.13	0.03	0.07
(**FCC**) (**valeur approchée**) F_i	0.17	0.50	0.77	0.90	0.93	1
(**FCD**) (**valeur approchée**) F_i	1	0.83	0.50	0.23	0.1	0.07

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

1- Déterminer les paramètres caractéristiques. (Pour une variable quantitative discrète)

Pour une classe de 30 élèves, on connait le nombre de frères et sœurs de chaque élève. D'aprés le calcule 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 x_i	0	1	2	3	4	5
Effectif n_i	5	10	8	4	1	2
ECC N_i	5	15	23	27	28	30
ECD N_i	30	25	15	7	3	2
Fréquence f_i (valeur approchée)	0.17	0.33	0.27	0.13	0.03	0.07
(**FCC**) (**valeur approchée**) F_i	0.17	0.50	0.77	0.90	0.93	1
(**FCD**) (**valeur approchée**) F_i	1	0.83	0.50	0.23	0.1	0.07

TAB. 2.3 – Tableau de nombre de fréres et soeurs d'une classe

1- Déterminer les paramètres caractéristiques.

1- **Les paramètres caractéristiques :** 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 $(N = 30)$,$\left(\frac{N}{2} = \frac{30}{2} = 15\right)$. alors :

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

$M_e = 1.5$. Cela signifie que la médiane n'est pas nécessairement une des valeures 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 :

$$\overline{X} = \sum_{i=1}^{6} f_i x_i = [(0.17\times0)+(0.3\times1)+(0.3\times2)+(0.13\times3)+(0.03\times4)+(0.07\times5)] \simeq 1.73.$$

- On remarque

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

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

– **Les paramètres de dispersion :**

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

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

$$E = 5.$$

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

– *Les quartiles :*

Le 1^{er} quartile : on a $N = 30, \frac{30}{4} = 7.5$.

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

- $Q_1 = x_{1/4} = x_2 = 1$. Cela signifie que 25% des élèves ont 1 frère et sœur.

- $Q_2 = M_e = 1.5$.

- Le $3^{iéme}$ quartile : $3 \times \frac{N}{4} = 3 \times \frac{30}{4} = 22.5$.

- $Q_3 = x_{3/4} = x_3 = 2$. (75% des des élèves ont 2 frères et sœurs.).

– *L'intervalle interquartile :*

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

– *Les déciles :*

- Le 1^{er} décile : on a $\frac{N}{10} = \frac{30}{10} = 3$.

$$D_1 = x_{1/10} = x_1 = 0.$$

- Le $5^{iéme}$ décile : $D_5 = M_e = 1.5$.

- Le $9^{iéme}$ décile : on a $9 \times \frac{N}{10} = 9 \times \frac{30}{10} = 27$.

$$D_9 = x_{9/10} = x_4 = 3.$$

– *L'intervalle interdécile :*

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

– *Les centiles :*

- Le 1^{er} centile : on a $\frac{N}{100} = \frac{30}{100} = 0.3$.

$$C_1 = x_{1/100} = x_1 = 0.$$

- Le $50^{iéme}$ centile : $C_{50} = M_e = 1.5$.

- Le $99^{iéme}$ centile : on a $99 \times \frac{N}{100} = 99 \times \frac{30}{100} = 29.7$.

$$C_{99} = x_{99/100} = x_6 = 5.$$

– *L'intervalle intercentile :*

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

L'écart absolu moyen :

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

– *La variance :* on va la calculer d'aprés 2.22

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

- On peut aussi la calculer à partir de 2.25

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

– *L'écart-type :* on a :

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

– *Coefficient de variation :*

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

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

– **Les paramètres de forme :**

– *Coefficient d'asymétrie de Pearson :*

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

– *Coefficient d'asymétrie de Yule :*

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

– *Coefficient d'asymétrie de Fisher :*

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

On remarque que tous les coefficients d'asymétrie sont supérieurs ou égaux à 0, et comme $\overline{X} > M_e > M_o$ ce qui singnifie 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.

Chapitre 3

Application sous R

3.1 Exemple sur la variable qualitative :

Exemple 3.1.1 *(Variable qualitative ordinale).1.5 Notre exemple est basé sur on interroge 50 personnes sur leur dernier diplôme obtenu (variable Y) La codification a été faite selon le Tableau 1.1. On a obtenu la série*

Dernier diplôme obtenu	X_j
Sans diplôme	Sd
Primaire	P
Secondaire	S
Supérieur non-universitaire	Su
Universitaire	U

TAB. 3.1 – Codification de la variable Y

1). Tracer le diagramme en barres des effectifs cumulés et diagramme en secteurs des fréquences

Lecture des données :YY=c("Sd","Sd","Sd","Sd","P","P","P","P","P","P","P","P","P","P","P

YF=factor(YY,levels=c("Sd","P","Se","Su","U"))

$\mathbf{x}_j$	$\mathbf{n}_j$	$\mathbf{N}_j$	$\mathbf{f}_j$	$\mathbf{F}_j$
Sd	4	4	0.08	0.08
P	11	15	0.22	0.30
Se	14	29	0.28	0.58
Su	9	38	0.18	0.76
U	12	50	0.24	1.00
	50		1.00	

TAB. 3.2 – Tableau statistique complet

T2=table(YF)

V2=c(T2)

data.frame(Eff=V2,EffCum=cumsum(V2),Freq=V2/sum(V2),

FreqCum=cumsum(V2/sum(V2)))

2. La représentation graphique : Instruction R.

- Pour tracer le diagramme en barres des effectifs cumulés : On utilise la fonction barplot(y).

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

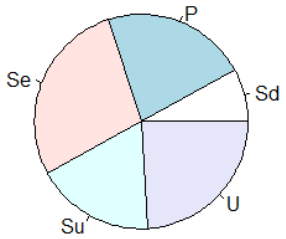


FIG. 3.1 – Diagramme en secteurs des fréquences

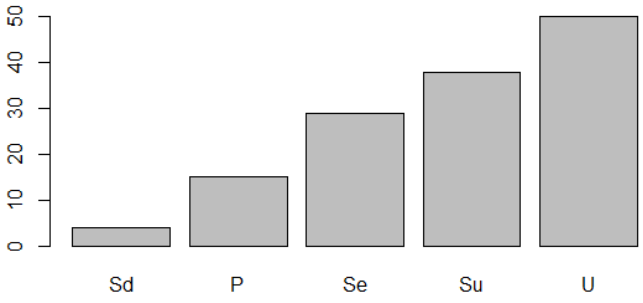


FIG. 3.2 – Diagramme en barres des effectifs cumulés

3.2 Exemples sur la variable quantitative :

3.2.1 Cas d'une variable quantitative discrète :

Exemple 3.2.1 *Dans une petite localité, on a relevé le nombre de pièces par appartement :*

Nombre de pièces	*1*	*2*	*3*	*4*	*5*	*6*	*7*	***Total***
Nombre d'appartements	*48*	*72*	*96*	*64*	*39*	*25*	*3*	*347*

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

Lecture des données :

> X=c(1,2,3,4,5,6,7) # Nombre de pièces.

> Y=c(48,72,96,64,39,25,3) # Nombre d'appartements.

> 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	3	3.175793

Les paramètres de dispersion :

Les paramètres	Code **R**	Les résultats
L'étendue	$\max(D) - \min(D)$	6
La variance	$\text{var}\,(D)$	2.156869
L'écart-type	$\text{sd}\,(D)$	1.468628
Le cofficient de variation	$\text{sd}(D)/\,\text{mean}(D)$	0.4624447
L'écart absolu moyen	$\text{mean}(\text{abs}(X - \text{mean}(D)))$	1.83203
Intervalle interquartile	$\text{IQR}(D)$	2

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

Les résultats :

0%	25%	50%	75%	100%
1	2	3	4	7

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)).

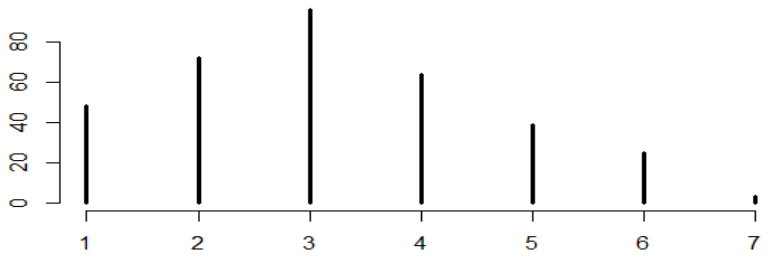


FIG. 3.3 – Digaramme en batôn

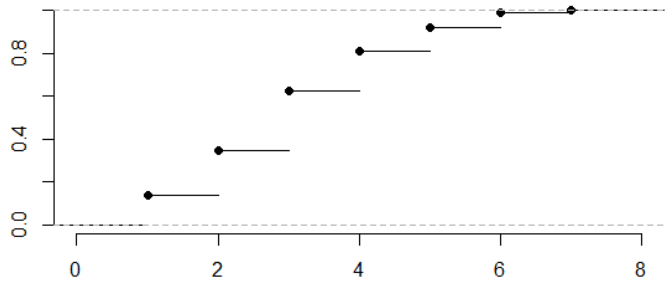


FIG. 3.4 – Diagramme cumulatif

3.2.2 Cas d'une variable quantitative continue

Exemple 3.2.2 *Pour 20 élèves des classes différentes, on connait les résultats d'un examan d'Mathmetique, on obtient le tableau suivant :*

$\mathbf{X}_i$	$\mathbf{n}_i$	$\mathbf{a}_i$	$\mathbf{c}_i$	$\mathbf{N}_i$	f_i	$\mathbf{n}_i'$
$[0;2[$	2	1	2	2	0.1	1
$[2;5[$	2	3.5	3	4	0.1	0.66
$[5;8[$	3	6.5	3	7	0.15	1
$[8;10[$	4	9	2	11	0.2	2
$[10;14[$	5	12	4	16	0.25	1.25
$[14;16[$	3	15	2	19	0.15	1.5
$[16;20[$	1	18	4	20	0.05	0.25
Total	20				1	

1. *Déterminer les paramètres caractéristiques.*
2. *Tracer l'histogramme des fréquences corrigées.*

Le logiciel R nous donne les résultats suivants :*Les paramètres caractéristiques :*

Les paramètres de tendance centrale :

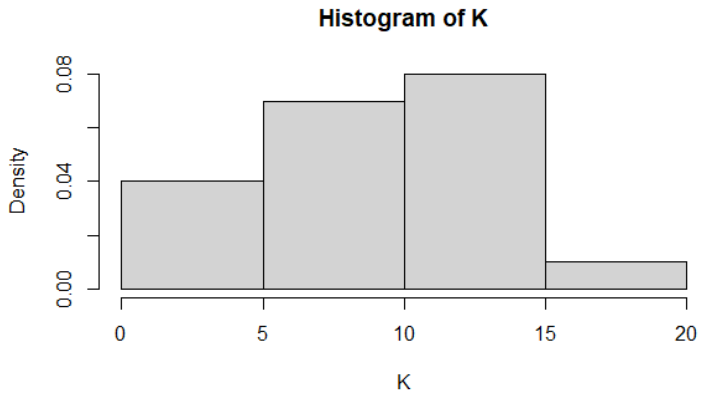
Le mode	*La mediane*	*La moyenne*
5	9	9.375

Les paramètres de dispersion :

Les paramètres	*Code* ***R***	*Les résultats*
L'étendue	$\max(K) - \min(K)$	17
La variance	$\text{var}\,(K)$	23.07566
L'écart-type	$\text{sd}(K)$	4.80
Le cofficient de variation	$\text{sd}(K)/\,\text{mean}(K)$	0.51
L'écart absolu moyen	$\text{mean}(\text{abs}(Ci - \text{mean}(K)))$	6.73
Intervalle interquartile	$IQR(K)$	5.5

La représentation graphique : Instruction ***R*** *.*

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



(a)

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.

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

Bibliographie

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

[2] Betteka. S., (2018 − 2019). Statistique descriptive univariée.

[3] Chekroun. A. (2017 − 2018). Statistiques descriptives et exercices.

[4] Diouri, M., Elmarhoum, A.(2006). Statistiques descriptives Cours et exercices.

[5] Hammdani, H. (2001). Statistique descriptive. Office des publications Universitaires : 02 − 2001.

[6] Immediato, H. (2001). Licence Scientifique. Cours. Statistiques.

[7] Khechai, I.,(2021). Statistique descriptive univariée .

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

[9] Sellam, M.,(2018). Statistique descriptive univariée.

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

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

[12] Sayah Abdallah.(2022). Cours de statistique descriptive univariée .

Notations et symbols

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
$\text{card}\,(\Omega)$	Le cardinal : nombre d'éléments de l'ensemble Ω .
f_i	Fréquence partielle
$f_i^{'}$	La fréquence corrigée
$n_i^{'}$	L'effectif corrigé
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_α^p	Moment d'ordre p
ECD	L'effectif cumulé décroissant
FCD	fréquence cumulée décroissante
ECC	L'effectif cumulé croissant
FCC	fréquence cumulée croissante
P	Le coefficient de Pearson
F	Le coefficient de Fisher
β_2	Coefficient d'aplatissement de Pearson
F_2	Coefficient d'aplatissement de Fisher.
EAM	L'écart absolu moyen

Abstract

Univariate descriptive statistics is the study of data associated with a single variable, whether it is a qualitative or quantitative variable, which is interested in summarizing the phenomenon studied with two techniques:

The graphical representation: gives an overall shape on the data distribution and simplifies the results, and the characteristic parameters (position, dispersion, shape): which gives interpretations of the results obtained.

Keywords: Descriptive statistics, statistical variable, quantitative, qualitative, graphical representation, characteristic parameters: position, dispersion, shape.

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.

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

ملخص

الإحصاء الوصفي أحادي المتغير هو دراسة البيانات المرتبطة بمتغير واحد سواء كان متغيرًا نوعيًا أو كميًا ، ويهتم بتلخيص الظاهرة المدروسة بتقنيتين:

التمثيل الرسومي: يعطي شكلاً شاملاً لتوزيع البيانات ويبسط النتائج ، والمعلمات المميزة (الموضع ، والتشتت ، والشكل): التي تعطي تفسيرات للنتائج التي تم الحصول عليها.

الكلمات المفتاحية: الإحصاء الوصفي، المتغير الإحصائي، الكمي، النوعي، التمثيل البياني، المعلمات المميزة: الموضع، التشتت، الشكل.