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

ZEMMOUCHE Hanane

Titre :

La statistique descriptive

Membres du Comité d'Examen :

Pr.	CHERFAUOI Mouloud	UMKB	Président
Pr.	SAYAH Abdallah	UMKB	Encadreur
Dr.	ABDELLI Jihane	UMKB	Examinateur (rice)

Juin 2024

Dédicace

Je dédie ce humble travail à :

Mes chers parents

Tous les membres de ma famille

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

Moral.

A mon encadreur, Monsieur

Sayah Abdallah

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

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 :

Pr. SAYAH ABDALLAH

Et lui souhaite le meilleur dans sa carrière professionnelle.

Je tiens à remercier :

Pr. CHERFAUOI Mouloud, **Dr. ABDELLI Jihane**

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

Je remercie tous ce qui a contribué dans ce travail,

De prés ou de loin.

Table des matières

Table des figures

Liste des tableaux

Introduction

La statistique est l'une des importantes branches en mathématiques, avec diverses applications et ce sont d'élément essentiel dans chaque thèse scientifique.

Etude des statistiques, également appelée statistique, est une méthode scientifique utilisée pour collecter, analyser et traiter un ensemble de données afin d'organiser et d'analyser les résultats, fournissant ainsi la base pour prendre des décisions.

Elle décrit les phénomènes en se basant sur des représentations graphiques telles que les diagrammes et les diagrammes cumulatifs, ...etc. On peut utilisée des mesures de tendance centrale comme la médiane, la moyenne et le mode peuvent être utilisés pour comprendre les valeurs centrales de cette variable. De plus, l'écart type et la variance peuvent être utilisés pour comprendre la distribution des valeurs et leur dispersion. Ces méthodes nous aident à obtenir une idée initiale sur les données et leurs caractéristiques générales

Et cette méthode est la première utilisée en statistique. Les statistiques se divisent en deux catégories :

• **Statistique descriptive** : La statistique descriptive est un ensemble de méthodes permettant de décrire, présenter, résumer des données souvent très nombreuses.

• **Statistique inférentielle** : La statistique inférentielle est d'effectuer des estimations et des prévisions à partir d'un sous-ensemble de population.

Dans notre étude, nous nous intéresserons aux statistiques descriptives, qui à leur tour, sont composées en deux catégories.

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

2 **La statistique descriptive multivariée** : Les analyses multivariées, c'est l'étude de la population à plusieurs variables. Les statistiques descriptives bi variées sont des cas particuliers à deux variables.

Les questions qui se posent à travers l'étude du sujet des statistiques descriptives, sont nombreuses, parmi elles, par exemple :

Qu'est-ce que les statistiques descriptives ?

Quel est l'intérêt des statistiques descriptives dans notre vie quotidienne ?

Quelles sont les étapes des statistiques descriptives ?

Dans cette thèse scientifique, nous nous concentrerons sur quatre chapitres pour l'étude de la statistique descriptive :

Chapitre 1 : Généralités sur la statistique

Ce chapitre est consacré à la définition de certains termes et concepts fondamentaux statistique, ainsi qu'à la connaissance des types de variables. Il se termine par la définition de la série statistique.

Chapitre 2 : Statistique descriptive univariée

Ce chapitre présente l'étude des séries statistiques avec une seule variable et nous souhaitons les relier aux paramètres. Ces derniers permettent de collecter (ou réduire) les informations, et les trois paramètres principaux sont lesètres de position, de dispersion et de forme.

Chapitre 3 : Statistique descriptive bivariée

Ce chapitre vise à décrire et analyser une série statistique avec deux variables, en mettant en évidence la relation entre elles. Parmi les outils les plus importants figurent le diagramme de dispersion, le coefficient de corrélation et la régression linéaire. Le but est de comprend de manière globale l'interaction entre les variables, et d'aider à prendre des décisions éclairées basées sur les données.

Chapitre 4 : Application par le logiciel R

Nous allons présenter quelques exemples concernants les trois chapitres précédents en utilisant le logiciel de programmation R

Chapitre 1

Généralités sur la statistique

La statistique est un groupe de méthodes scientifiques qui permettent de collecter et d'organiser des données ou des observations numériques, pour étudier un sujet. It est nécessaire d'utiliser certains concepts et terminologies de la statistique car ils sont considères comme les mots clés pour accéder à cette science , et on retrouve parmi eux : échantillon , population , individu ....

Dans ce premier chapitre, nous abordons la définition de quelques concepts de base utilisés en statistique descriptive

1.1 Définitions fondamentales

La 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.

La statistique descriptive

La statistique descriptive est une branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données. Cette démarche a pour but de :

• Résumer et synthétiser l'information contenue dans la série statistique.

• Mettre en évidence ses propriétés.

• Suggérer des hypothèses relatives à la population dont est issu l'échantillon.

Les Outils utilisés sont :

• Les Tableaux.

• Les Graphiques.

• Les indicateurs.

Le type d'outils utilisés dépendent de :

1. La nature de la série (univariée ou multidimensionnelle).
2. La nature des variables (quantitatives ou qualitatives).

Si les données ne sont relatives qu'à une seule variable, on parle de statistique descriptive " univariée".

Dans le cas ou l'on s'intéresse à deux variables simultanément, on met en œuvre la statistique descriptive " bivariée ". Si l'ensemble de données provient de l'observation de plusieurs variables, on doit faire appel aux méthodes de la statistique descriptive " multivariée "

1.2 Terminologie de base statistique

On précise ici un certain nombre de termes statistiques très courants qui seront régulièrement utilisés par la suite et qu'il convient de bien connaître :

La population :

La population est un ensemble d'individus définis par une propriété commune donnée.si l'on veut étudier la durée de vie des ampoules électriques fabriquées par une compagnie, la population considérée est l'ensemble de toutes les ampoules fabriquées par cette compagnie.

Exemple : Si nous voulons étudier le degré d'utilisation du téléphone, la population serait l'ensemble de toutes les personnes abonnées à celui-ci

Individu :

On appelle individu chaque élément de la population ou de l'échantillon (sous ensemble ou partie de la population). On utilise également le terme (unité statistique) pour désigner un individu.

Exemple : Dans l'exemple précédent portant sur le degré d'utilisation du téléphone, chaque personne abonnée représente un individu ou une unité statistique.

Les modalités :

Les modalités sont les différentes situations dans lesquelles les individus peuvent se trouver.Â l'égard du caractère considéré.

Exemple : Le sexe est un caractère qui présente deux modalités : féminin ou masculin.

Echantillon :

L'échantillon est un sous ensemble de la population considérée. Le nombre d'individus dans l'échantillon est la taille de l'échantillon. pour établir la durée de vie des ampoules électriques produites par une machine, on peut

prélever au hasard un certain nombre d'ampoules - un échantillon- parmi toutes les celles produites par cette machine.

1.3 Caractère ou variable statistique

Chaque individu d'une population est décrit par un ensemble de caractéristiques appelées variables ou caractères, on les représente souvent par des lettre majuscules : X,Y,....

Les valeurs qui peut prendre une variable statistique sont appelées modalités. Une variable doit donc présenter au minimum deux modalités.

Une variable statistique peut être soit qualitative ou quantitative.

1.3.1 Variable statistique qualitative

Définition :

Un caractère est dit qualitatif s'il n'est pas mesurable.

Exemple de caractères qualitatifs :

La couleur de cheveux d'un groupe de personnes, le sexe des étudiants d'une école (masculin ou féminin), état civil des individus (marié, célibataire, divorcé) etc...

Une variable statistique qualitative peut être nominale ou ordinale

• **Variable qualitative nominale** : La variable est dite qualitative nominale quand les modalités ne peuvent pas être ordonnées

• **Variable qualitative ordinale** : La variable est dite qualitative ordinale quand les modalités peuvent être ordonnées. Le fait de pouvoir ou non ordonner les modalités est parfois discutable. Par exemple : dans les catégories socioprofessionnelles, on admet d'ordonner les modalités : 'ouvriers', 'employés', 'cadres'. Si on ajoute les modalités 'sans profession', 'enseignant', 'artisan', l'ordre devient beaucoup plus discutable..

1.3.2 Variable statistique quantitative

Une variable est dite quantitative si toute ses valeurs possibles sont numériques.

La variable statistique quantitative peut être discrète ou continue

• **Variable quantitative discrète** : Une variable est dite discrète, si l'ensemble des valeurs possibles est dénombrable.

Les variables quantitatives discrète sont des caractères dont les modalités sont des nombres isolés, pas nécessairement entiers les observations d'une variable quantitative discrète sont en général des ($\in \mathbb{N}$). Dans certains cas, elles peuvent être des décimales.

Nombre de pièces d'un immeuble.

Nombre d'enfants d'une famille.

Nombre de personnes atteintes par une maladie.

• **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.

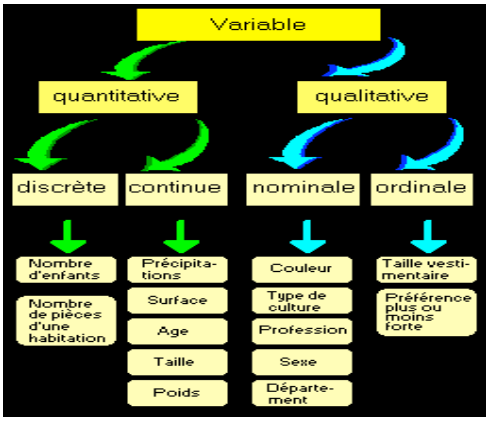


FIG. 1.1 – Types de variable

1.4 Types de série statistique

1.4.1 Série statistique univariée

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

Le nombre d'unités d'observation est noté p. Les valeurs de la variable X sont notées

$$x_1, x_2, ..., x_p.$$

1.4.2 Série statistique bivariée

On s'intéresse à deux variables x et y ces deux variables sont mesurées sur les n unités d'observation. Pour chaque unité on obtient donc deux mesures. La série statistique est alors une suite de n couples des valeurs prises par les deux variables sur chaque individu :

$$(x_1, y_1), ..., (x_i, y_j), ..., (x_n, y_n)$$

Chacune des deux variables peut être, soit quantitative, soit qualitative.

Notation

On notera x_i , $i = 1, ..., k$ les modalités ou valeurs de la variable de X

On notera y_j , $j = 1, ..., m$ les modalités ou valeurs de la variable de Y

Les deux variables X et Y sont mesurées simultanément sur chacun des N individus de la population. On notera n_{ij} l'éffectif correspondant au couple (x_i, y_j)

1.5 Notations

Voici les terminologies et notations usuelles pour les définitions ci-dessous :

Terminologie	Notation
Taille de population	N
Population[2]	$\mathcal{P} = \{1, ..., N\}$
Individu	$u \in \mathcal{P}$
Variable	$X, Y,$
Donnée de variable X pour l'individu u	$X(u)$
Série statisique (simple) brute pour X	$\{X(1), ..., X(N)\}$
Série statisique(double) brute pour X et Y	$\{(X(1), Y(1)), ..., (X(N), Y(N))\}$

TAB. 1.1: Notation

Exemple 1.5.1

Supposons qu'on va faire une étude sur le sexe, l'âge, la taille, la masse ... des étudiants inscrits à la $1^{ère}$ année de l'université de biskra.

1. Identifier la population, les individus, les variables statistiques et leurs modalités.

Population : Tout étudiants inscrits à la $1^{ère}$ année de l'université de Biskra.

Individu : Tout étudiant inscrit à la $1^{ère}$ année.

Caractères

$X =$ sexe

$Y =$ âge(ans)

$U =$ taille(cm)

$V =$ masse(Kg)

Modalités :

$$X = \{\text{masculin, féminin}\}$$

$$Y = \{18 \text{ ans}, ..., 30 \text{ ans}\}$$

$$U = [130, 210]$$

$$V = [40, 120]$$

Chapitre 2

Statistique descriptive univariée

La statistique descriptive univarée consiste à décrire chaque lettre statistique une à une et non les liens entre elles il est utilisé pour afficher les données observées .pour une variable sous forme de tableaux et de graphiques et les résumer numériquement avec des indicateurs statistiques. Toutes ces méthodes nous aident à facilement des séries statistiques une variée.

2.1 Effectifs et fréquences

2.1.1 Effectif

L'effectif de la valeur x_i , $1 \leq i \leq p$ d'un caractère est le nombre d'individus de la population ayant cette valeur, elle est notée n_i .

Effectif total

L'effectif total est le nombre de données, la somme des effectifs des modalité x_i on le note N.

$$N = \sum_{i=1}^{p} n_i$$

2.1.2 Fréquence

La fréquence d'une valeur x_i du caractère notée f_i est le quotient de l'effectif de la valeur par l' effectif total c'est- à-dire f_i :

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

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.

• Les fréquences sont souvent données en pourcentage : on multiplie alors chaque résultat par 100, par **exemple** si $f_1 = 0.24$. Ainsi $0.24 \times 100 = 24$, alors on dit qu'on a 24% de la population ont la modalité1 Alors on a :

$$\sum_{i=1}^{p} f_i = 1$$

ou

$$\sum_{i=1}^{p} f_i = 100$$

Dans le cas des fréquences en pourcentage.

Effectif cumulé et fréquence cumulée

L'effectif cumulé croissant (la fréquence cumulée croissante) d'une modalité x_i est la somme de tous les effectifs (de toutes les fréquences) jusqu'au rang i compris. Ils sont notés respectivement N_i et F_i ou (ECC et FCC)

$$N_i = n_1 + n_2 + ... + n_i \quad \text{et} \quad F_i = f_1 + f_2 + ... + f_i$$

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

$$N_i = n_p + n_{p-1} + ... + n_i \quad \text{et} \quad F_i = f_p + f_{p-1} + ... + f_i.$$

Remarque 2.1.1

Fréquence absolue = effectif

Fréquence relative = fréquence

Fréquence cumulée = fréquence relative cumulée

2.2 Tableau statistique en génerale

Un tableau statistique constitue un résumé ou une synthèse numérique des résultats d'une distribution statistique, on distingue trois formes de tableaux statistiques qui sont fonction de l'objectif envisagé et de la nature du caractère étudié.

2.2.1 Cas d'un caractère quantitatif continu :

Le tableau statistique :

Les classe C_i	center d'un classe	n_i	f_i	FCC
$[b_1; b_2[$	c_1	n_1	f_1	F_1
$[b_2; b_3[$	c_2	n_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[b_{p-1}; b_p[$	c_p	n_p	f_p	F_p= 1
$\sum$		N	1	

TAB. 2.1 – Tableau statistique regroupé par classes

On commence par notations suivantes :

Série ordonnée : les valeurs obtenues peuvent être rangées par ordre de grandeur par exemple croissante. On obtient une série statistique ordonnée.

Etendue de la série : la différence entre les deux valeurs extrêmes est appelée étendue de la série.

Classe : quand le caractère étudié est quantitatif continu, la série statistique est répartie en classes ou intervalles semi ouverts. Le nombre de classes, k est calculé par l'une des deux formules :

La règle de Sturge$k = 1 + 3.3log(N)$

La règle de Yule $k = 2.5(N)^{1/4}$

Soit la classe de la modalité $C_i = [b_{i-1}, b_i[$, on note de manière générale :

$c_i = \dfrac{b_i + b_{i-1}}{2}$ le centre de la classe C_i,

b_{i-1} la borne inférieure de classe C_i

b_i la borne supérieure de classesC_i

$a_i = b_i - b_{i-1}$ l'amplitude de la classe C_i

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

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

2.2.2 Cas d'une variable qualitative et quantitive discréte :

Le tableau statistique permet de résumer la série statistique en faisant un regroupement des individus associés aux modalités auxquelles ils appartiennent :

Les modalités X_i	n_i	ECC	f_i	FCC
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	$N_p = 1$	f_p	$F_p = 1$
$\sum$	N		1	

TAB. 2.2 – Tableau statistique d'un caractére qualitatif et quantitatif discret

2.3 Les variables qualitatives

Un caractère est dit qualitatif lorsque ses modalités ne sont pas mesurables. Le nombre de valeurs que peut prendre la variable est limité. Il existe au sein de ce type deux échelles : nominale et ordinale.

2.3.1 Echelle nominale

Chaque modalité est exprimée par un nom ou un code. Les différentes modalités ne sont pas ordonnables.

Exemple 2.3.1 *: Cas des noms*

Etat matrimoniale : marié, célibataire, veuf, divorcé.

Sexe : féminin, masculin.

Profession : enseignant, médecin.

Nationalité : Algérienne, Tunisienne.

Les différentes séquences nucléotidiques.

Les hormones : œstradiol, progestérone.

Exemple 2.3.2 *: Cas des codes*

Etat matrimoniale : marié (1), célibataire (2), veuf (3), divorcé (4).

Sexe : féminin (1), masculin (2).

Profession : enseignant (1), médecin (2).

Nationalité : Algérienne (1), Tunisienne (2).

2.3.2 Echelle ordinale

Chaque modalité est explicitement significative du rang pris par chaque individu pour le caractère considéré.

Degré d'intelligence : pas intelligent (0), peu intelligent (1), moyennement intelligent (2), très intelligent (3).

Forme des fruits : petite (1), moyenne (2), grosse (3).

Abondance/Dominance : peu abondant (1), abondant (2), très abondant (3).

2.3.3 Tableau statistique de variable qualitative

Un tableau statistique constitue un résumé ou une synthèse numérique des résultats d'une distribution statistique.

Cas d'une variable qualitative nominale :

Exemple 01

La série suivante représente la spécialité choisie par 100 étudiants de première année, le tableau statistique :

Choix des étudiants X_i	n_i	f_i
Mathématique	57	$57/100 = 0.57$
Economie	33	$33/100 = 0.33$
Informatique	10	$10/100 = 0.1$
$\sum$	100	1

Cas d'une variable qualitative ordinale :

Exemple 02

On interroge 50 personnes sur leur dernier diplôme obtenu (variable Y). La codification a été faite selon. On a obtenu la série 1 Dernier diplôme obtenu X_i : Sans diplôme Sd, Primaire P, Secondaire S, Supérieur non-universitaire Su, Universitaire U

La série statistique est :

Sd, Sd, Sd, Sd, P, P, P, P, P, P, P, P, P, P, P, Se, Se,Se, Se, Se, Se, Se, Se, Se, Se, Se, Se, Se, Se, Su, Su, Su, Su, Su, Su, Su, Su, Su, U, U, U, U, U, U, U, U, U, U, U, U

La tableaux statistique est :

X_i	n_i	N_i	f_i	F_i
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
$\sum$	50		1	

2.3.4 Représentations graphiques

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont :

1. **Diagramme en secteurs (Diagramme circulaire) :** Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence.

Le degré d'un secteur est déterminé à l'aide de la règle de trois de la manière suivante :

$$N \rightarrow 360°$$
$$n_i \rightarrow b_i \text{ (degrés de la modalité).}$$

Donc pour les effectifs : $b_i = \frac{n_i}{N} \times 360°$

et pour les fréquences : $b_i = f_i \times 360°$

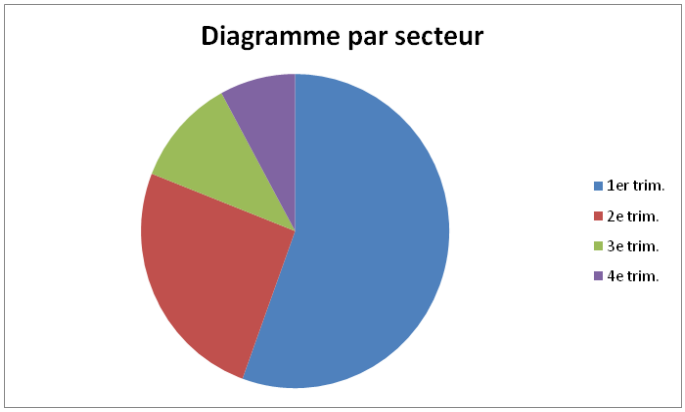


FIG. 2.1 – Diagramme par secteur

2. **Diagramme en bâtonnet :** Le diagramme en bâtonnets (ou tuyaux d'orgue) est une représentation graphique de la distribution de fréquences d'une variable qualitative. Les "bâtonnets" sont bien séparés pour indiquer les déférents Catégories. La hauteur d'un batonnet est proportionnelle à la fréquence de la catégorie correspondante
3. **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.

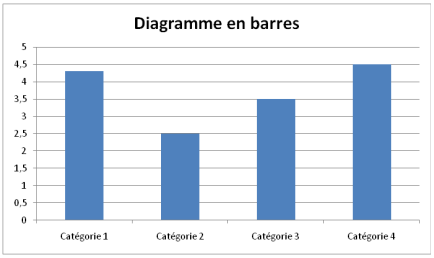


FIG. 2.2 – Diagramme en barres

Exemple 02 d'une variable qualitative ordinale est :

Présentation diagramme en secteur :

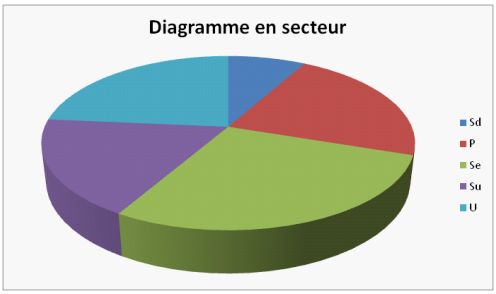


Diagramme en secteur des effectifs

Présentation diagramme en barres :

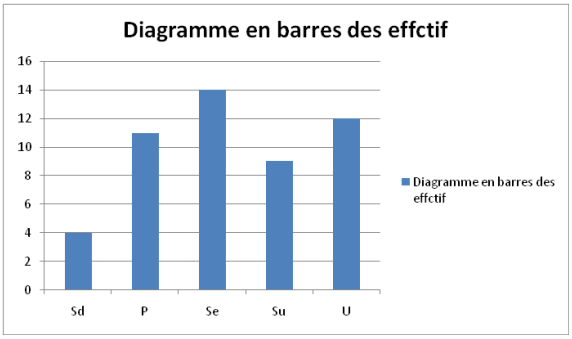


FIG. 2.3 – Diagramme en barres des effectifs

2.4 Les variables quantitatives

Un caractère est dit quantitatif si ses modalités s'expriment par des nombres dont les opérations de types sommes et produits sont possibles sur les valeurs des modalités. Le nombre de valeurs que peut prendre la variable est illimité. On distingue deux catégories de caractère quantitative :

2.4.1 Tableau statistique d'une variable quantitative

Cas d'une variable quantitative discrète

Exemple03

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

Modalité x_i	0	1	2	3	4	5	6
Effectif 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	$0.05 + 0.19 = 0.24$	$100 - 5 = 95$	$95 \div 100 = 0.95$
2	27	51	0.27	$0.24 + 0.27 = 0.51$	$95 - 19 = 76$	$76 \div 100 = 0.76$
3	21	72	0.21	$0.51 + 0.21 = 0.72$	$76 - 27 = 49$	$49 \div 100 = 0.49$
4	15	87	0.15	$0.72 + 0.15 = 0.87$	$49 - 21 = 28$	$28 \div 100 = 0.28$
5	9	96	0.09	$0.87 + 0.09 = 0.96$	$28 - 15 = 13$	$13 \div 100 = 0.13$
6	4	100	0.04	$0.96 + 0.04 = 1$	$13 - 9 = 4$	$4 \div 100 = 0.04$
Tatal	100		1			

Exemple 04

Dans une petite localité, on a relevé de nombre de pièces par appartement

nombre de pièces X_i	1	2	3	4	5	6	7
nombre de d'appartement n_i	48	72	96	64	39	25	3

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	f_i	F_i
1	48	48	0.139	0.139
2	72	120	0.207	0.346
3	96	216	0.277	0.623
4	64	280	0.184	0.807
5	39	319	0.112	0.919
6	25	344	0.072	0.991
7	3	347	0.009	1

Cas d'une variable quantitative continu

Exemple 05

Le taux de glucose sanguin (glycémie) déterminé chez 32 sujets est donné ci-dessous en g/l

Série ordonnée :

0.85 0.87 0.90 0.93 0.94 0.94 0.95 0.97 0.97 0.98 0.98 0.99 1 1.01 1.03 1.03 1.03 1.04 1.06 1.07 1.08 1.08 1.10 1.10 1.11 1.13 1.14 1.14 1.15 1.17 1.19 1.21

1. Déterminer le tableau statistique

On a $N = 32$

Nombre de classes par deux formules :

La règle de Sturge$k = 1 + 3.3log(N) = 1 + 3.3\log(32) = 5.966 \simeq 6$

La règle de Yule $k = 2.5(N)^{1/4} = k = 2.5(32)^{1/4} = 5.946 \simeq 6$

Etendue de série $1.21 - 0.85 = 0.35$

Amplitude d'une classe $a_i = 0.06$

Tableaux de statistique :

Classée g/l	n_i	c_i	N_i= ECC	f_i	F_i= FCC	d_i	δ_i
$[0.85, 0.91[$	3	0.88	3	$3/32 = 0.09$	0.09	50	1.5
$[0.91, 0.97[$	4	0.94	$3 + 4 = 7$	$4/32 = 0.13$	$0.09 + 0.13 = 0.22$	66.67	2.17
$[0.97, 1.03[$	7	1.00	$7 + 7 = 14$	$7/32 = 0.22$	$0.22 + 0.22 = 0.44$	116.67	3.67
$[1.03, 1.09[$	8	1.06	$8 + 14 = 22$	$8/32 = 0.25$	$0.25 + 0.44 = 0.69$	133.33	4.17
$[1.09, 1.15[$	6	1.12	$6 + 22 = 28$	$6/32 = 0.18$	$0.69 + 0.18 = 0.87$	100	3
$[1.15, 1.21]$	4	1.18	$4 + 28 = 32$	$4/32 = 0.13$	$0.13 + 0.87 = 1$	66.67	2.17
$\sum$	32						

2.4.2 Représentation graphique les variables quantitatives

Cas de variables discrètes :

Dans le cas des séries statistiques discrètes il existe deux types de représentations graphiques :

La représentation en Diagramme en bâtons : est la représentation de la distribution des fréquences ou des effectifs d'une variable discrète. A chaque valeur x_i portée en abscisse on fait correspondre un segment vertical de longueur proportionnelle à l'effectif ni ou à la fréquence f_i de cette valeur.

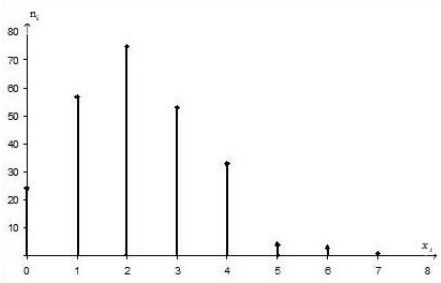


FIG. 2.4 – Diagramme en batôn

La courbe cumulative : est la représentation graphique des effectifs cumulés ou des fréquences cumulées. C'est un graphique en escalier dont les paliers horizontaux ont pour ordonnées respectivement f_i ou N_i les marches de l'escalier correspondent aux valeurs possibles x_i de la variable statistique x et sont à des hauteurs proportionnelles aux effectifs cumulés ou aux fréquences cumulées.

• La courbe cumulative est la représentation graphique de la proportion $F(x)$ des individus de la population pour lesquels la valeur de la variable statistique est inférieure ou égale à x. Cette fonction, définie pour toute valeur de x, est appelée fonction cumulative ou fonction de répartition. Elle est constante dans chaque intervalle séparant deux valeurs de la variable statistique :$F(x) = \begin{cases} 0 & x < x_1 \\ F_i & x_i < x < x_{i+1} \\ 1 & x_i < x \end{cases}$

On peut aussi définir la fonction de répartition de la variable statistique x, notée aussi $F(x)$, comme la ligne brisée qui joint les milieux des paliers de la courbe cumulative.

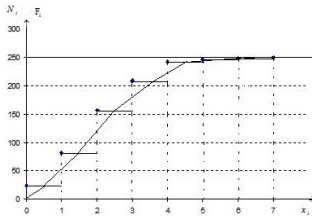
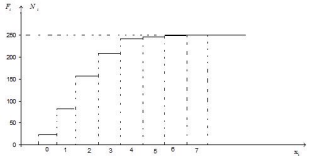


FIG. 2.5 – Fonction de répartition



Cas de variables continues :

L'histogramme : L'histogramme est la représentation graphique de la distribution des effectifs ou des fréquences d'une variable statistique continue. A chaque classe de valeurs de la variable statistique portée en abscisse, on

fait correspondre un rectangle basé sur cette classe. Alors chaque modalité est représentée par un rectangle dont l'aire (et non la hauteur) est proportionnelle à la fréquence ou à l'effectif de cette classe.

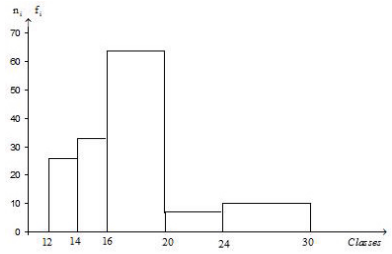
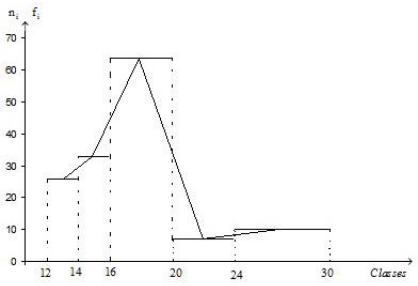


Diagramme en histogramme

· La courbe des fréquences est la fonction en escalier dont les paliers sont constitués par les bases supérieures des rectangles formant l'histogramme des fréquences.

· Le polygône des fréquences est la ligne brisée qui relie les milieux des cotés supérieurs des rectangles de l'histogramme des fréquences.



Courbe cumulative

La courbe cumulative est la représontation graphique des effectifs cumulés ou des fréquences cumulées C'est un graphique en escalier dont les paliers horizontaux ont pour ordonnées respectivement F_i ou N_i Comme pour les variables discrètes, la courbe cumulative ou histogramme des fréquences cumulées, est la représentation graphique de la fonction cumulative ou fonction de répartition $F(x)$ est une fonction de $\mathbb{R}$ dans $[0,1]$, qui est définie par :

$$F(x) = \begin{cases} 0 \text{ si } x < b_1 \\ F_{i-1} + \dfrac{f_i}{b_{i+1} - b_i}\ (x - b_i) \text{ si } b_i \leq x < b_{i+1} \\ 1 \text{ si } x_i \leq b_p \end{cases}$$

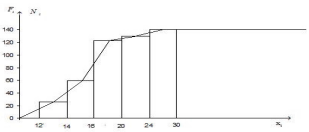


FIG. 2.6 – Courbe cumulative

2.4.3 Paramétres de position

Moyenne :

Cas d'une variable discrète

Soit $x_1, ..., x_p$ les valeurs du caractère , $n_1, ..., n_p$ les effectifs correspondants et $N = n_1 + ... + n_p$ l'effectif total. La moyenne de la série statistique est le nombre :

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

• La définition précédente est celle de la moyenne arithmétique. On définit aussi d'autres moyennes pour $x_i > 0$ comme :

–la moyenne harmonique h telle que : $h = \frac{1}{N} \sum_{i=1}^{p} \frac{n_i}{x_i}$

–la moyenne géométrique g telle que : $g = (x_1^{n_1} \times ... \times x_p^{n_p})^{\frac{1}{N}}$

Cas d'une variable continue

La quantité est : $\overline{x} = \sum_{i=1}^{k} f_i c_i$.

S'appelle la moyenne de X

c_i : centre de classes

f_i : fréquence partielle de la classe

Médiane :

Cas d'une variable discrète

Soit $x_1, ..., x_p$ les valeurs de la variable statistiques classées par ordres croissant ,on distingue alors deux cas :

N est impair: $N = 2k + 1$ alors : $M_e = x_{k+1}$

N est pair: $N = 2k$ alors : $M_e = \frac{x_k + x_{k+1}}{2}$

Cas d'une variable continue

Pour calculer la médiane dans le cas continu, on dit doit tout d'abord déterminer la classe médiane qui correspond soit à :

La valeur de $N_i \uparrow$ supérieur ou égale à $N/2$

La valeur de $f_i \uparrow$supérieur ou égale à0.5

On calcule la médiane par cette formule :

$$M_e = b_i + \frac{N/2 - N_{i-1}}{n_i} \times a_i$$

Dans le cas où nous utilisons les fréquences relatives cumulées croissantes, le calcul de la médiane se fait, par la formule suivent :

$$M_e = b_i + \frac{0.5 - F_{i-1}}{f_i} \times a_i$$

Tel que :

M_e : La valeur de la médiane.

b_i : La borne inférieur de la classe médiane.

N_{i-1} : L'effectif cumulé croissant de la classe si classe située avant la classe médiane.

n_i : La valeur de l'effectif de la classe médiane.

a_i : L'amplitude de la classe .

N : Le'effectif total de population.

Détermination la médian graphiqment

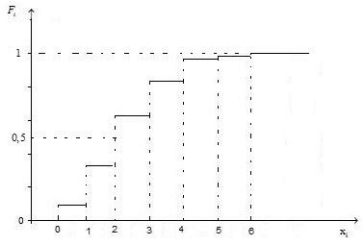


FIG. 2.7 – Détermination graphique de la médiane : variable discrète

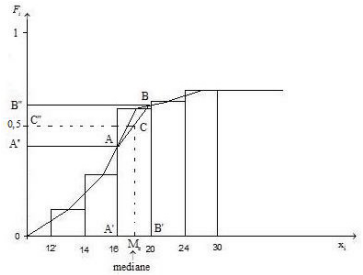


FIG. 2.8 – Détermination graphique de la médiane : variable continue

Le mode

Cas d'une variable discrète

Le mode d'une $V.S$ est la valeur qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par M_o.

Cas d'une variable continue

Nous définissions la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif partiel (ou la plus grande fréquence partielle).

La quantité :

$$M_o = b_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times a_i$$

S'appelle le mode :

b_i : la borne inférieure de la classe modale.

a_i : le pas de la classe modale.

$\Delta_1 = n_0 - n_1,\ \Delta_2 = n_0 - n_2$

$\Delta_1 = f_0 - f_1,\ \Delta_2 = f_0 - f_2$

n_0 et f_0 sont l'effectif et la fréquence associés à la classe modale.

n_1 et f_1 sont l'effectif et la fréquence de la classe qui précède la classe modale.

n_2 et f_2 sont l'effectif et la fréquence de la classe qui suit la classe modale

Détermination le mode graphique

• Lorsque la variable est discrète le mode est défini avec précision. Si, par exemple, deux valeurs successives de la variable statistique ont la même fréquence maximum, on dit qu'il y a un intervalle modal dont les extrémités correspondent à ces deux valeurs.

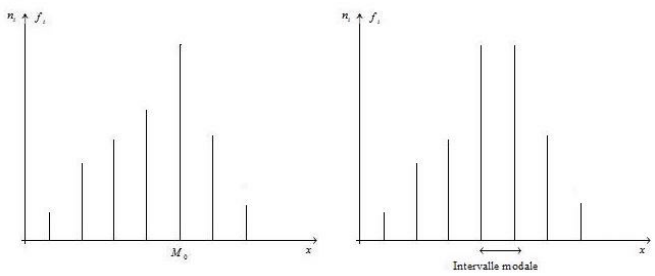


FIG. 2.9 – Détermination le mode graphiquement

• Lorsque la variable est continue, la détermination du mode est beaucoup moins précise car les fréquences

dépendent du découpage en classe. L'utilisation de la courbe des fréquences ajustée sur l'histogramme, bien que peu précise, conduit à une bonne estimation du mode dans le cas où les classes sont d'égale amplitude.Il est possible d'établir une formule d'interpolation linéaire pour le calcul exacte du mode dans le cas d'une répartition en classes d'amplitude quelconque

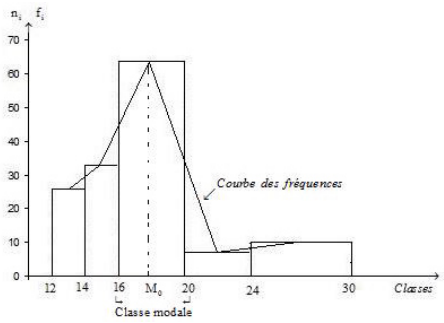


FIG. 2.10 – Détermination Mode graphiquement

Exemple 2.4.1

Déterminer paramètres de position

Dans l'exemple 05 :

$\overline{x} = 0.88 \times 0.09 + 0.94 \times 0.13 + 1 \times 0.22 + 1.06 \times 0.25 + 1.12 \times 0.18 + 0.18 \times 0.13 = 1.04$

$M_o = 1.03 + \frac{8-7}{(8-7)+(8-6)} \times (1.09 - 1.03) = 1.05$

On a $N \times 0.5 = 16$

$M_e = 1.03 + \frac{16-14}{8} \times (1.09 - 1.03) = 1.045$

Pour exemple 03

$\overline{x} = \sum_{i=1}^{p} f_i x_i = 0 \times 0.05 + 1 \times 0.19 + 2 \times 0.27 + 3 \times 0.21 + 4 \times 0.15 + 5 \times 0.09 + 6 \times 0.04 = 2.65$

On a $N = 100$ est pair alors $M_e = \frac{x_{50}+x_{51}}{2} = \frac{2+2}{2} = 2$

Précédent, le mode est égal à 2 qui correspondant au plus grand effectif. c.-à-d. est $M_o = 2$

Pour exemple 04

$\overline{x} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i = 3.17579250$

On a.$N = 347$ est impair alors : $M_e = x_{173+1} = x_{147} = 3$

Précédent, le mode est égal à 2 qui correspondant au plus grand effectif. c.-à-d. est $M_o = 2$

2.4.4 Paramètres de dispersion

Les indicateurs statistiques de dispersion usuels sont l'étendue, la variance et l'écart-type.

L'étendue

La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité

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

• S'appelle l'étendue de la $V.S$ X. Le calcul de l'étendue est très simple. Il donne une première idée de la dispersion des observations

Quantile

Cas d'une variable discrète

La notion de quantile d'ordre p (où $0 < p < 1$) généralise la médiane. Formellement un quantile est donné par l'inverse de la fonction de répartition :

$x_p = F^{-1}(p)$

Si la fonction de répartition était continue et strictement croissante, la définition du quantile serait sans équivoque La fonction de répartition est cependant discontinue et "par palier". Quand la fonction de répartition est par palier, il existe au moins 9 manières différentes de définir les quantiles selon que l'on fasse ou non une interpolation de la fonction de répartition. Nous présentons une de ces méthodes, mais il ne faut pas s'étonner de voir les valeurs des quantiles différer légèrement d'un logiciel statistique à l'autre.

— Si Np est un nombre entier, alors

$$x_p = \tfrac{1}{2}\{x_{(Np)} + x_{(Np+1)}\}$$

— Si Np n'est pas un nombre entier, alors

$$x_p = x_{(\lceil Np \rceil)}$$

Où $\lceil Np \rceil$ représente le plus petit nombre entier supérieur ou égal à Np.

— La médiane est le quantile d'ordre $p = 1/2$.

— On utilise souvent :

$x_{1/4}$ le premier quartile.

$x_{3/4}$ le troisième quartile.

$x_{1/10}$ le premier décile.

$x_{1/5}$ le premier quintile.

$x_{4/5}$ le quatrième quintile.

$x_{9/10}$ le neuvième décile.

$x_{0.05}$ le cinquième percentile.

$x_{0.95}$ le nonante-cinquième percentile.

— Si $F(x)$est la fonction de répartition, alors $F(x) \geq p$

• **Le quartile** : qui divise l'effectif total de la distribution statistique en 4 parties égales (notée $Q_1,\ Q_2,\ Q_3$)

• **Le décile** : qui divise l'effectif total de la distribution statistique en 10 parties égales (notée $D_1\ , ...,\ D_9$)

• **Le percentile** : qui divise l'effectif total de la distribution statistique en 100 parties égales (notée $C_1\ , ...,\ C_{99}$)

Cas d'une variable continue

Les quantiles se calculent de la même manière que la médiane :

$$q_\alpha = b_i + \frac{N \times \alpha - N_{i-1}}{n_i} \times a_i$$

Tel que : α est le paramètre de quantile c-à-d :$(Q_1 : \alpha = 1/4), (Q_3 : \alpha = 3/4), (D_1 : \alpha = 1/10, (D_9 : \alpha = 9/10), (C_1 : \alpha = 1/100, C_{99} : \alpha = 99/100)$

La boîte à moustaches

Il s'agit d'un diagramme permettant de positionner les quartiles Q_1, Q_2, Q_3, au moyen de rectangles de largeur arbitraire, prolongés par des "moustaches" de part et d'autre, de longueur au plus égale à une fois et demie $Q_3 - Q_1$.

Ces diagrammes sont surtout utiles pour comparer rapidement l'allure générale de plusieurs distributions.

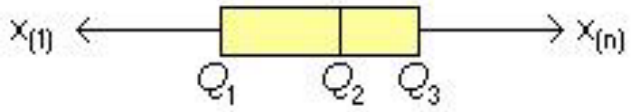


FIG. 2.11 – La boite à moustaches

Les intervalles interquantiles

Le quartile, le décile et le centile se détermine de la même manière que la médiane :

a L'intervalle interquartile : est la différence entre les valeurs du troisième et du premier quartile $IQ = (Q_3 - Q_1)$ L'intervalle $[Q_1, Q_3]$ contient 50% des valeurs de X. Plus IQ est grand, plus X est dispersée.

b L'intervalle interdécile : est la différence entre les valeurs du neuvième et du premier décile $ID = (D_9 - D_1)$ L'intervalle $[D_1, Q_3]$ contient 80% des observations.

c L'intervalle intercentile : est la différence entre les valeurs du quatre vingt dix neuvième et du premier centile $IC = (C_{99} - C_1)$ L'intervalle $[C_1, C_{99}]$ contient 98% des observations.

La variance

Cas d'une variable discrète

On appelle variance de cette série statistique X, le nombre

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

On dit que la variance est la moyenne des carrés des écarts à la moyenne x. Les « écarts àla moyenne » sont les $(\overline{x} - x_i)$, les « carrés des écarts à la moyenne » sont donc les $(\overline{x} - x_i)^2$. En faisant la moyenne de ces écarts, on trouve la variance.

• Le théorème suivant (Théorème de König-Huygens) donne une identité remarquable reliant la variance et la moyenne, parfois plus pratique dans le calcule de la variance.

Soit (x_i, n_i) une série statistique de moyenne $\overline{x}$ et de variance $Var(X)$. Alors.

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

Cas d'une variable continue

La variance est la quantité :

$$Var(X) = \sum_{i)1}^{k} f_i(\overline{x} - c_i)^2.$$

Pour le calcul, on utilise théorème

$$Var(x) = \sum_{i)1}^{k} f_i {c_i}^2 - \overline{x}^2$$

L'écart type

L'écart type : est la racine carrée de variance : $\sigma_x = \sqrt{Var(X)}$

Le coefficient de variation

Lorsqu'on veut comparer la dispersion ou l'étalement de deux séries d'observations qui n'ont pas le même ordre de grandeur ou qui portent sur des variables différentes, on ne peut pas utiliser directement les écarts types. Le coefficient de variation se définit comme le rapport de l'écart type divisé par la moyenne, exprimé en pourcentage.

$$CV = \frac{\sigma_x}{\overline{x}}$$

Exemple 2.4.2

Déterminer paramètres de dispersion

Dans l'exemple 05 : on a

Les quartiles :

On a : $N \times 0.25 = 8$ alors $Q_1 = 0.97 + \frac{8-14}{8} \times (1.09 - 1.03) = 1.045$

On a : $N \times 0.75 = 24$ alors $Q_3 = 1.11$

Les déciles :

ona $N \times 0.1 = 3.2$ alors $D_1 = 0.913$

ona $N \times 0.9 = 28.8$ alors $D_9 = 1.162$

Les intervalles interquantile est :

$IQ = (Q_3 - Q_1) = 1.11 - 1.045 = 0.065$

$ID = (D_9 - D_1). = 1.162 - 0.913 = 0.249$

La variance : $Var(X) = \sum_{i=1}^{k} f_i(\overline{x} - C_i)^2 = 0.008$

L'écart type est : $\sigma_x = \sqrt{Var(X)} = 0.089$

Le coefficient de variation est : $CV = \frac{\sigma_x}{\overline{x}} = \frac{0.089}{1.04} = 0.085$

Dans l'exemple 03 :

Précédent $E = x_{\max} - x_{\min} = 6 - 0 = 6$

On a $N = 100$

Le premier quartile comme $Np = 100 \times 0.25 = 25$ est un nombre entier on a $Q_1 = x_{1/4} = \frac{x_{25}+x_{26}}{2} = 2$

Le troisième quartile comme $Np = 100 \times 0.75 = 75$ est un nombre entier on a $Q_3 = x_{3/4} = \frac{x_{75}+x_{76}}{2} = 4$

Le premier décile comme $Np = 100 \times 0.1 = 10$ est un nombre entier on a $D_1 = x_{1/10} = \frac{x_{10}+x_{11}}{2} = 1$

Le décile comme $Np = 100 \times 0.9 = 90$ est un nombre entier on a $D_9 = x_{9/10} = \frac{x_{90}+x_{91}}{2} = 5$

Le premier centile comme $Np = 100 \times 0.01 = 1$ est un nombre entier on a $C_1 = x_{1/100} = \frac{x_1+x_2}{2} = 0$

Le centile comme $Np = 100 \times 0.99 = 99$ est un nombre entier on a $C_{99} = x_{99/100} = \frac{x_{99}+x_{100}}{2} = 6$

Les intervalles interquantile est :

$IQ = (Q_3 - Q_1) = 4 - 2 = 2$

$ID = (D_9 - D_1). = 5 - 1 = 4$

$IC = (C_{99} - C_1) = 6 - 0 = 6$

$Var(X) = \sum_{i=1}^{p} f_i x_i^2 - \overline{x}^2 = 2.2275$

L'ecart type est : $\sigma_x = \sqrt{Var(X)} = 1.4925$

Le coefficient de variation est : $CV = \frac{\sigma_x}{\overline{x}} = \frac{1.4925}{2.65} = 0.56321$

donc $CV = 0.56321 > 0.5$, alors la dispersion est importante

Dans l'exemple 04 :

Précédent $E = x_{\max} - x_{\min} = 7 - 1 = 6$

On a $N = 347$

Le premier quartile comme $Np = 347 \times 0.25 = 86.75$ n'est pas un nombre entier, alors $Q_1 = x_{1/4} = 2$

Le troisième quartile comme $Np = 347 \times 0.75 = 260.25$ n'est pas un nombre entier, alors $Q_3 = x_{3/4} = 4$

Le premier décile comme $Np = 347 \times 0.1 = 34.7$ n'est pas un nombre entier, alors $D_1 = x_{1/10} = 1$

Le décile comme $Np = 347 \times 0.9 = 312.3$ n'est pas un nombre entier, alors $D_9 = x_{9/10} = 5$

Le premier centile comme $Np = 347 \times 0.01 = 3.47$ n'est pas un nombre entier, alors $C_1 = x_{1/100} = 1$

Le centile comme $Np = 347 \times 0.99 = 343.53$ n'est pas un nombre entier, alors $C_{99} = x_{99/100} = 6$

Les intervalles interquantile est :

$IQ = (Q_3 - Q_1) = 4 - 2 = 2$

$ID = (D_9 - D_1) = 5 - 1 = 4$

$IC = (C_{99} - C_1) = 6 - 1 = 5$

$Var(X) = \sum_{i=1}^{p} f_i x_i^2 - \overline{x}^2 = 2.15065319$

L'ecart type est : $\sigma_x = \sqrt{Var(X)} = 1.46651054$

Le coefficient de variation est : $CV = \frac{\sigma_x}{\overline{x}} = \frac{1.46651054}{3.17579250} = 0.4618$

L'écart moyen absolu : L'écart moyen absolu est la somme des valeurs absolues des écarts à la moyenne divisée par le nombre d'observations :

$$e_{moy} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \overline{x}|$$

L'écart médian absolu : L'écart médian absolu est la somme des valeurs absolues des écarts à la médiane divisée par le nombre d'observations :

$$e_{med} = \frac{1}{N} \sum_{i=1}^{N} \left|x_i - x_{1/2}\right|$$

Moment

Pour $r \in \mathbb{N}$ on définit le moment d'ordre r : $m_r = \frac{1}{N} \sum_{i=1}^{p} n_i x_i^r = \sum_{i=1}^{p} f_i x_i^r$

Le moment centré d'ordre r :$\mu_r = \frac{1}{N} \sum_{i=1}^{p} n_i \left(x_i - \overline{X}\right)^r = \sum_{i=1}^{p} f_i \left(x_i - \overline{X}\right)^r$

$\mu_2 = m_2 - (m_1)^2$

$\mu_3 = m_3 - 3m_1m_2 + 2(m_1)^3$

$\mu_4 = m_4 - 4m_1m_3 + 6(m_1)^2 m_2 - 3(m_1)^4$.

μ_3 et μ_4 sont utilisés dans le calcul des paramètres de forme.

2.4.5 Les paramatres de forme

Ces paramètres permettent de préciser la forme de la distribution expérimentale. Ils affinent la description de la distribution d'une variable et facilite la comparaison de plusieurs

Distributions expérimentales. Les paramètres de forme que nous aborderons sont :

L'asymétre : L'asymétrie d'une distribution peut être approchée par une comparaison entre le mode, la médiane et la moyenne. On dit qu'une distribution est symétrique si les trois valeurs de tendance centrale sont égales :Moyenne=Médiane=Mode.

Ainsi, le premier quartile et le troisième quartile, le premier décile et le neuvième décile. Le premier centile et le 99^c centile, sont équidistance à la médiane :

$$Q_3 - M_e = M_e - Q_1$$

$$D_9 - M_e = M_e - D_1$$

$$C_{99} - M_e = M_e - C_1$$

Les principaux coefficients d'asymétrie sont valables que si les série sont uni modales et si la variable statistique prend pour un nombre assez élevé de valeurs.

Le coefficient d'asymétrie : Il permet de nous renseigner sur la façon régulière ou non dont les observations se répartissent de part et d'autre d'une valeur centrale.

1. **Le coefficient d'symétrie Fisher** : c'est la quantité suivante

$$F = \frac{\mu_3}{\sigma_x^3}$$

Il s'agit du moment à l'origine d'ordre 3 des données centrées réduites.

2. **Le coefficient d'asymétrie de Yule :** est basé sur les positions des 3 quartiles ($1er$ quartile, médiane et $3eme$ quartile),et est normalisé par la distance interquartile :

$$Y = \frac{Q_1+Q_3-2Q_2}{Q_3-Q_1}$$

3. **Le coefficient de Pearson** : est basé sur une comparaison de la moyenne et du mode, et est standardisé par l'´ecart-type :

$$P = \frac{\overline{x}-M_o}{\sigma_x}.$$

Remarque 2.4.1

Tous les coefficients d'asymétrie ont les mêmes propriétés, ils sont nuls si la distribution est symétrique, négatifs si la distribution est allongée à gauche (left asymmetry), et positifs si la distribution est allongée à droite (right asymmetry).

Dans exemple 04 :

Le coefficient de Fisher $F = 0.35 > 0$

Le coefficient de Yule $Y = 0.14 > 0$

Le coefficient de Pearson $P = 0.435\,51 > 0$

Donc la distribution est étalée à droite.

L'aplatissement :

Le coefficient d'aplatissement : L'aplatissement est mesuré par **le coefficient d'aplatissement de Pearson :**

$$\beta_2 = \frac{\mu_4}{\sigma_x^4}$$

Ou **le coefficient d'aplatissement de Fisher :**

$$F_2 = \beta_2 - 3 = \frac{\mu_4}{\sigma_x^4} - 3$$

- Une courbe mésokurtique si.$F_2 \approx 0$
- Une courbe leptokurtique si $F_2 > 0$. Elle est plus pointue et possède des queues plus longues.
- Une courbe platykurtique si $F_2 < 0$. Elle est plus arrondie et possède des queues plus courtes

L'coefficient d'aplatissement de kelley : il existe un autre coefficient d'aplatissement défini à l'aide des quartiles et les déciles :.

$$C_k = \frac{1}{2} \times \frac{Q_3 - Q_1}{D_9 - D_1}$$

Chapitre 3

Statistique descriptive bivariée

Nous avons vu dans le chapitre précédent que les statistiques descriptives univariées se concentrent, pour une population donnée, sue caractéristique particulière. Nous pouvons maintenant étudier, visualiser et mesurer les liens existants entre deux variables pour la même population. Dans ce chapitre, nous nous intéressons à l'étude simultanée de deux variables étudiées sur la même population. Ce type de statistiques est utilisé pour comprendre la relation et mesurer le niveau de corrélation entre elles, c'est pourquoi nous aborderons l'étude des statistiques descriptive deux variables.

3.1 Effectifs et Fréquences

3.1.1 Effectifs

Effectifs joints

On définit l'effectif du couple (x_i, y_j) comme le nombre n_{ij} des données tel que $X = x_i$ et $Y = y_j$

$$n_{ij} = Card\{u \in \Omega \ / X(u) = x_i \ , Y(u) = y_j\}$$

Effectifs marginales

Effectifs marginale par rapport à Y : nous avons, pour $j = 1...m$,

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

Effectifs marginale par rapport à X : nous avons, pour $i = 1...k$,

$$n_{i\bullet} = n_{i1} + n_{i2} + ...n_{im} = \sum_{j=1}^{m} n_{ij}$$

3.1.2 Fréquences

Fréquences jointes

On appelle fréquence de l'évènement $(x_i\ ,y_j)$ la proportion des observations qui présentent simultanément les modalités x_i et y_j . Elle est notée f_{ij} et est définie telle que : $f_{ij} = n_{ij}/N$.

Fréquences marginales

Il est évident que d'après la définition de la fréquence f_i,On a :

$$f_{i\bullet} = \frac{n_{i\bullet}}{N} \quad i = 1, 2, ..., k \text{ et } f_{\bullet j} = \frac{n_{\bullet j}}{N} \quad j = 1, 2, ..., m$$

Définition

Distribution marginale des fréquences par rapport à X le total des fréquences de la ligne i.

$$f_{i\bullet} = \sum_{j=1}^{m} f_{ij} = \sum_{j=1}^{m} \frac{n_{ij}}{N} = \frac{n_{i\bullet}}{N}$$

et la fréquence marginale par rapport à Y le total des fréquences de la colonne j.

$$f_{\bullet j} = \sum_{i=1}^{k} f_{ij} = \sum_{i=1}^{k} \frac{n_{ij}}{N} = \frac{n_{\bullet j}}{N}$$

Remarque 3.1.1

$$\sum_{i=1}^{k} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{k} f_{i\bullet} = \sum_{j=1}^{m} f_{\bullet j} = 1$$

3.2 Tableaux statistique à deux variable

Tableau des effectifs

Considérons un échantillon comprenant N individus, alors les valeurs prises par les deux variables X et Y sont respectivement :

$$\begin{aligned} X &= (x_1, x_2, ..., x_k) \\ Y &= (y_1, y_2, ..., y_m)\,. \end{aligned}$$

Les deux variables ayant respectivement k et m modalités.

La distribution des effectifs est l'application qui, à chaque couple de valeurs (x_i, y_j) associé son effectif partiel n_{ij}.

Définition

On appelle tableau de contingence, ou tableau à double entrée, où figurent en colonnes les modalités de X et en lignes les modalités de Y. Le nombre d'éléments du sous ensemble est l'effectif n_{ij} des individus présentant à la fois les modalités x_i du caractère X et y_j du caractère Y.

$X \diagdown Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{km}	$n_{k\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet_j}$	...	$n_{\bullet_m}$	N

TAB. 3.1 – Tableau effectif dans le cas bivarié

Pour pouvoir lire les éléments contenus dans le tableau de contingence, des conventions de notations ont été établies.

Les modalités x_i de la variable X apparaissent en colonne et les effectifs n_{ij} apparaissent en ligne i.

Les modalités y_j de la variable Y apparaissent en ligne et les effectifs n_{ij} apparaissent en colonne j.

Tableau des fréquences

Le tableau de fréquences s'obtient en divisant tous les effectifs par la taille de l'échantillon :

$X \diagdown Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	Total
x_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1m}	$f_{1\bullet}$
x_2	n_{21}	n_{22}	...	f_{2j}	...	f_{2m}	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	...	f_{ij}	...	f_{im}	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	f_{k1}	f_{k2}	...	f_{kj}	...	f_{km}	$f_{k\bullet}$
total	$f_{\bullet 1}$	$f_{\bullet 2}$	...	$f_{\bullet_j}$	...	$f_{\bullet_m}$	1

TAB. 3.2 – Tableau fréquence dans le cas bivarié

Exemple

On a effectué une enquête sur 10 individus en observant le poids X en kg et la taille Y en centimètre

Les résultats sont donnés dans le tableau suivant :

X	65	60	64	71	74	76	78	80	82	84
Y	159	161	158	168	170	169	175	178	181	186

1) Déterminer le tableau de contingence (X : poids, Y : taille). Pour le poids et pour la taille, former respectivement des classes de pas de 5 kg et de 4 cm.

2) Déterminer le tableau statistique des deux séries marginales X et Y.

1) Le tableau des effectifs

$X \diagdown Y$	$[158, 162[$	$[162, 166[$	$[166, 170[$	$[170, 174[$	$[174, 178[$	$[178, 182[$	$[182, 186]$	$n_{i\bullet}$
$[60, 65[$	2	0	0	0	0	0	0	2
$[65, 70[$	1	0	0	0	0	0	0	1
$[70, 75[$	0	0	1	1	0	0	0	2
$[75, 80[$	0	0	1	1	0	0	0	2
$[80, 85[$	0	0	0	0	0	2	1	3
$n_{\bullet j}$	3	0	2	2	0	2	1	10

Ona $n_{1\bullet} = \sum_{j=1}^{7} n_{1j} = n_{11} + n_{12} + ... + n_{17} = 2$

$n_{2\bullet} = \sum_{j=1}^{7} n_{2j} = n_{21} + n_{22} + ... + n_{27} = 1$

La tableau des fréquence

$X \diagdown Y$	$[158, 162[$	$[162, 166[$	$[166, 170[$	$[170, 174[$	$[174, 178[$	$[178, 182[$	$[182, 186]$	$n_{i\bullet}$
$[60, 65[$	0.2	0	0	0	0	0	0	0.2
$[65, 70[$	0.1	0	0	0	0	0	0	0.1
$[70, 75[$	0	0	0.1	0.1	0	0	0	0.2
$[75, 80[$	0	0	0.1	0.1	0	0	0	0.2
$[80, 85[$	0	0	0	0	0	0.2	0.1	0.3
$n_{\bullet j}$	0.3	0	0.2	0.2	0	0.2	0.1	1

Ona : $f_{12} = 0$, $f_{21} = 0.1$ et $f_{43} = 0.1$

$f_{1\bullet} = \sum_{j=1}^{7} f_{1j} = \frac{n_{1\bullet}}{N} = \frac{2}{10} = 0.2$, et $f_{5\bullet} = \sum_{j=1}^{7} f_{5j} = \frac{n_{5\bullet}}{N} = \frac{3}{10} = 0.3$

2) Le tableau statistique de la série marginale de X et Y

X	$n_{i\bullet}$	$f_{i\bullet}$	Centre de la classe
$[60, 65[$	2	0.2	62.5
$[65, 70[$	1	0.1	67.5.
$[70, 75[$	2	0.2	72.5
$[75, 80[$	2	0.2	77.5
$[80, 85[$	3	0.3	82.5

et

Y	$n_{\bullet j}$	$f_{\bullet j}$	Centre de la classe
$[158, 162[$	3	0.3	160
$[162, 166[$	0	0	164
$[166, 170[$	2	0.2	168
$[170, 174[$	2	0.2	172
$[174, 178[$	0	0	176
$[178, 182[$	1	0.1	180
$[182, 186[$	2	0.2	184

3.3 Représentation graphique de distribution deux caractères

Le mode de représentation graphique d'une distribution à deux caractères n'est strictement possible que dans un espace à trois dimensions. Chacun des caractères est porté sur une dimension et la troisième est affectée aux effectifs ou aux fréquences.

3.3.1 Cas des caractères quantitatifs

Une série statistique double dont les caractères X et Y sont quantitatifs est représentée par les points M_i de coordonnées $(x_i y_j)$ dans un repère orthogonal du plan. Cette représentation s'appelle nuage de points de la série statistique double.

3.3.2 Cas des caractéres qualitatifs

Si les deux variable X et Y sont qualitatives, alors les données observées sont une suite de couples de variable (x_1, y_1),...,$(x_i$,$y_j)$,...,$(x_N$,$y_N)$ il n'est pas possible, dans ce cas, de représenter les deux caractères de façon absolument symétrique.

3.4 Descriptions numériques

3.4.1 Les moyennes marginales

Les moyennes marginales sur un échantillon de X et Y de taille N sont définies par :

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{m} n_{\bullet j} x_i = \sum_{i=1}^{k} f_{i\bullet} x_i \quad \text{et} \quad \overline{y} = \frac{1}{N}\sum_{j=1}^{m} n_{\bullet j} y_j = \sum_{j=1}^{m} f_{\bullet j} y_j$$

Dans le cas continu, x_i et y_i représentent respectivement le centre des classes de X et Y , c'est à dire :

$$x_i = \frac{L_{i+1}+L_i}{2} \quad \text{et} \quad y_j = \frac{L_{j+1}+L_j}{2}$$

3.4.2 Les variances marginales

Les variances marginales de X et Y sont définies par :

$$S_x^2 = \sigma_x^2 = Var\,(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}(x_i - \overline{x})^2 = \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i^2 - \overline{x}^2 = \sum_{i=1}^{k} f_{i\bullet}x_i^2 \ - (\overline{x})^2$$

et

$$S_y^2 = \sigma_y^2 = Var\,(Y) = \frac{1}{N}\sum_{j=1}^{m} n_{\bullet j}(x_j - \overline{x})^2 = \frac{1}{N}\sum_{j=1}^{m} n_{\bullet j}x_j^2 - \overline{x}^2 = \sum_{j=1}^{m} f_{\bullet j}y_j^2 - (\overline{y})^2$$

Les écarts-type de X et de Y sont donnés, respectivement,

$$\sigma_x = \sqrt{Var\,(X)} \text{ et } \sigma_y = \sqrt{Var\,(Y)}$$

Exemple 3.4.1

1) Calculons $\overline{x},\ Var\,(X)\,,\ \sigma_x$, et $\overline{y},\ Var\,(Y)\ ,\ \sigma_y$

$\overline{x} = \sum_{i=1}^{5} f_{i\bullet}x_i = 0.2 \times 62.5 + 0.1 \times 67.5 + 0.2 \times 72.5 + 0.2 \times 77.5 + 0.3 \times 82.5 = 74$

$$\begin{aligned} Var\,(X) &= \sum_{i=1}^{5} f_{i\bullet}x_i^2 - \left(\bar{X}\right)^2 \\ &= 0.2 \times (62.5)^2 + 0.1 \times (67.5)^2 + 0.2 \times (72.5)^2 + 0.2 \times (77.5)^2 + 0.3 \times (82.5)^2 - (74)^2 \\ &= 55.25 \end{aligned}$$

$\sigma_x = \sqrt{Var\,(X)} = \sqrt{55.25} = 7.433$

$\overline{y} = \sum_{j=1}^{7} f_{\bullet j}y_j = 0.3 \times 160 + 0.2 \times 168 + 0.2 \times 172 + 0.2 \times 180 + 0.1 \times 184 = 170.4$

$$\begin{aligned} Var\,(Y) &= \sum_{j=1}^{7} f_{\bullet j}y_j^2 \ - \left(\bar{Y}\right)^2 \\ &= 0.3 \times (160)^2 + 0.2 \times (168)^2 + 0.2 \times (172)^2 + 0.2 \times (180)^2 + 0.1 \times (184)^2 - (170.\,4) \\ &= 71.04 \end{aligned}$$

$\sigma_y = \sqrt{71.04} = 8.428\,5$

3.4.3 Covariance

En général, la distribution des observations d'une population suivant deux caractères $(x\ , y)$ sont disposées dans un tableau de contingence, alors la covariance est définie telle que :

Soit $(x\ , y)$ un couple de variables statistiques. On appelle covariance des variables statistiques x et y, notée $Cov(x\ , y)$, la quantité définie telle que :

$$Cov(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{m} f_{ij}\,(x_i - \overline{x})\,(y_j - \overline{y}) = \sum_{i=1}^{k}\sum_{j=1}^{m} f_{ij}x_i y_j - \overline{xy}$$

notation $f_{ij} = \frac{n_{ij}}{N}$

Dans certaines situations il arrive que les observations d'une population suivant deux caractères $(x\ , y)$ soient appariées, $i\ e.$ les observations sont disponibles sous forme d'une suite (x_i,y_i),$i = 1, 2, ..., N$, alors dans cette situation la covariance est définie telle que :

$$S_{xy} = Cov(X\ , Y) = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})\,(y_j - \overline{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_j - \overline{xy}$$

La covaraince indique le sens de la relation entre les variables X et Y ainis ,On peut distinguer les suivant :

si $Cov(x\ , y) > 0$,alors on peut dire que la relation entre les deux variables est positive . Dans ce cas, ces deux variables varient dans le même sens.

si $Cov(x\ , y) < 0$,alors on peut dire que la relation entre les deux variables est négative . Dans ce cas, ces deux variables varient en sens inverse .

si $Cov(x\ , y) = 0$,alors on peut dire qu'il n'y a pas de relation entre les deux variables . Dans ce cas, les variations de l'une n'entrainent pas la variation de l'autre .

3.4.4 Coefficient de corrélation linéaire.

On appelle coefficient de corrélation de deux variables statistiques$x$ et$y$, et on le note $Corr(x, y)$ ou , la quantité définie telle que :

$$\rho = Corr(x\ , y) = \frac{Cov(x\ , y)}{\sigma_x \sigma_y}$$

Ainisi, on peut distinguer les ces suivants :

Si $\rho > 0$, les deux variables varient dans le même sens

Si $\rho < 0$, les deux variables varient sens inverse.

Si $\rho = 0$, les deux variables sont linéairement indépendantes.

On a $-1 \leq \rho \leq 1$.

si $\rho = 1$ on dit qu'il y a une parfaite corréllation linéaire positive entre les deux variables

Si $\rho = -1$ on dit qu'il y a une parfaite corréllation linéaire négative entre les deux variables

Quelque soit le couple de variables statistiques $(x\ , y)$ leur coefficient de corrélation $\rho = Corr(x\ , y)$ vérifie l'inégalité suivante :

$-1 \leq \rho \leq 1$

Les égalités ont lieu si et seulement si il existe deux constantes $a \neq 0$ et b telles que $y = ax + b$ ou bien $x = ay + b$.

Exemple 3.4.2

Calculer

$$\begin{aligned} Cov(X,Y) &= \sum_{i=1}^{5}\sum_{j=1}^{7} f_{ij}x_i y_j - \bar{X}\bar{Y} \\ &= 0.1\,(67.5 \times 160 + 72.5 \times 168 + 72.5 \times 172 + 77.5 \times 168 + 77.5 \times 172 + 82.5 \times 184) \\ &+ 0.2\,(62.5 \times 160 + 82.5 \times 180) - 170.\,4 \times 74 \\ &= 58.4 \end{aligned}$$

Alors $\rho = corr(x\ , y) = \frac{58.4}{7.\,433 \times 8.428\,5} = 0.932\,18$

Comme 0.932 18 est proche de 1. Par conséquent, il existe une forte corrélation linéaire positive entre les deux caractères X et Y.

3.5 Ajustement linéaire ou droite des moindres carrés

On considère deux variables statistiques quantitatives X et Y et on s'intéresse à relation linéaire significative entre deux caractères

3.5.1 Droite de régression

La droite de régression est la droite qui ajuste au mieux un nuage de points au sens des moindres carrés.

On considère que la variable X est explicative et que la variable Y est dépendante. L'équation d'une droite est : $y = a + bx$

Le problème consiste à identifier une droite qui ajuste bien le nuage de points. Si les coefficients a et b étaient connus, on pourrait calculer les résidus de la régression définis par : $e_i = y_i - a - bx_i$

Le résidu e_i est l'erreur que l'on commet en utilisant la droite de régression pour prédire y_i à partir de x_i. Les résidus peuvent être positifs ou négatifs

3.5.2 Critère des moindres carrés

Pour déterminer la valeur des coefficients a et b on utilise le principe des moindres carrés qui consiste à chercher la droite qui minimise la somme des carrés des résidus :

$M(a,b) = \sum_{i=1}^{n} e_i^2 = \sum (y_i - a - bx_i)^2$

Les coefficients a et b qui minimisent le critère des moindres carrés sont donnés par

$b = \frac{S_{xy}}{S_x}$ et $a = \overline{y} - b\overline{x}$

Le minimum $M(a,b)$ en(a,b) s'obtient en annulant les dérivées partielles par rapport à a et b.

$$\frac{\partial M(a,b)}{\partial a} = -\sum_{i=1}^{n} 2(y_i - a - bx_i) = 0$$
$$\frac{\partial M(a,b)}{\partial b} = -\sum_{i=1}^{n} 2(y_i - a - bx_i)x_i = 0$$

On obtient un système de deux équations à deux inconnues. En divisant les deux équations par $-2n$, on obtient :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0$$

Ou encore

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} a - \frac{1}{n}\sum_{i=1}^{n} bx_i = 0$$
$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} ax_i - \frac{1}{n}\sum_{i=1}^{n} bx_i^2 = 0$$

Ce qui s'écrit aussi

$$\overline{y} = a + b\overline{x}$$
$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i - a\overline{x} - \frac{1}{n}\sum_{i=1}^{n} bx_i^2 = 0$$

La première équation montre que la droite passe par le point $(\overline{x}, \overline{y})$. On obtient

$a = \overline{y} - b\overline{x}$

En remplaçant a par $\overline{y} - b\overline{x}$ dans la seconde équation, reqm : $S_{xy} - bS_x^2 = 0$

donc

$b = \frac{S_{xy}}{S_x^2}$

On a donc identifié les paramètres

$$b = \frac{S_{xy}}{S_x^2}$$
$$a = \overline{y} - \frac{S_{xy}}{S_x^2}\overline{x}$$

La droite de régression est donc

$$y = a + bx = \overline{y} - \frac{S_{xy}}{S_x^2}\overline{x} + \frac{S_{xy}}{S_x^2}x$$

Ce qui peut s'écrire aussi

$$y - \overline{y} = \frac{S_{xy}}{S_x^2}(x - \overline{x})$$

La droite de régression de y en x n'est pas la même que la droite de régression de$x$ en y.

3.5.3 Résidus et valeurs ajustées

En x Les valeurs ajustées sont obtenues au moyen de la droite de régression : $\hat{y} = a + bx_i$

Les valeurs ajustées sont les 'prédictions' des yi réalisées au moyen de la variable x et de la droite de régression de y

3.5.4 Qualité d'Ajustement

Pour juger la qualité d'ajustement du modèle nous utilisons l'équation de l'analyse de la variance, c.-à-d. cherchons tout d'abord à décomposer la variance des y_i autour de leur moyenne en une somme de deux autre variance

.

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y})^2}_{SCR} + \underbrace{\sum_{i=1}^{n}(\hat{y} - \overline{y})^2}_{SCE}$$

Où

SCT : somme des carrés totale (ou variation totale des y_i).

SCR : somme des carrés résiduelle (ou variation des résidus e_i dite aussi variation résiduelle),

SCE : somme des carrés expliqués (variation expliquée).

La qualité d'ajustement peut être déterminée par le coefficient de détermination qui exprime le rapport entre la variation expliquée et la variation totale.

3.5.5 Coefficient de détermination.

On appelle coefficient de détermination, noté D, la quantité

$$D = R^2 = \frac{S_{xy}^2}{S_x S_y} = \frac{SCE}{SCT} < 1$$

Plus la variation résiduelle est proche de 0, c.-à.-d la variance expliquée est proche de la variance totale, plus R^2 proche de 1, plus l'ajustement est meilleur

Exemple

1. Déterminer l'équation de la droite de régression $\hat{y}_i = bx_i + a$
2. Estimer la taille d'une personne qui pèse 68 kg
3. Déterminer coefficient de détermination

1. On a : $b = \dfrac{58.4}{55.249} = 1.057$, et $a = 170.4 - 1.057 \times 74 = 92.182$, donc : $\hat{y}_i = 1.057x_i + 92.182$
2. $\hat{y} = 1.057 \times 68 + 92.182 = 164.06$, donc la taille d'une personne qui pèse $68kg$ est $164.06cm$
3. On a : $D = R^2 = \frac{S_{xy}^2}{S_x S_y} = \frac{58.4^2}{55.149 \times 71.04} = 0.868$

Chapitre 4

Application par le Logiciel R

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

4.1 Statistique descriptive univariée

4.1.1 Cas d'une variable qualitative

Exemple 4.1.1 *(variable qualitative nominale)*

On a pris un échantillon de 50 achats de boissons non-alcoolisées achetées dans une grande surface, en notant par : CC=Coca-Cola, S=Sprite, CL=Coke-Light, P=Perrier, PC=Pepsi-Cola. On a obtenu les résultats suivants :

Boisson	CC	CL	PC	P	S	Total
Effectif	19	8	13	5	5	50

1. Calculer les f_i , ECC ,FCC, le mode.
2. Tracer le diagramme en bandes et secteur angulaire.

Solution 4.1.1

Utiliser logiciel R

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

```
x=c(rep("CC",19),rep("CL",8),rep("PC",13),rep("P",5),rep("S",5))
y=c(table(x))
data.frame(Eff=y,ECC=cumsum(y),Frq=y/sum(y),FCC=cumsum(y/sum(y)))
which(y==max(y))
```

```
> x=c(rep("CC",19),rep("CL",8),rep("PC",13),rep("P",5),rep("S",5))
> y=c(table(x))
> data.frame(Eff=y,ECC=cumsum(y),Frq=y/sum(y),FCC=cumsum(y/sum(y)))
   Eff ECC  Frq  FCC
CC  19  19 0.38 0.38
CL   8  27 0.16 0.54
P    5  32 0.10 0.64
PC  13  45 0.26 0.90
S    5  50 0.10 1.00
> which(y==max(y))
CC
 1
```

FIG. 4.1 – les résultats en R

2 Tracer le diagramme en bandes et secteur angulaire

En langage R

x=c(rep("CC",19),rep("CL",8),rep("PC",13),rep("P",5),rep("S",5))

y=c(table(x))

pie(y,radius=1.0)#Diagramme en secteurs

m=max(y)

c=barplot(y,ylim=c(0,m+1))#Diagramme en barres

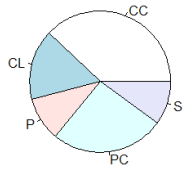


FIG. 4.2 – Diagramme en secteurs de l'effectif

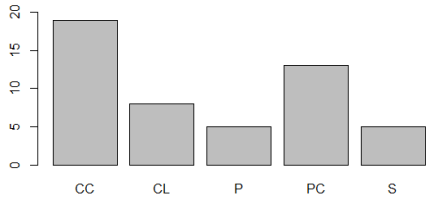


FIG. 4.3 – Diagramme en barres de l'effectif

4.1.2 Cas des variables quantitatives discrètes

Exemple 4.1.2

Un inspecteur en contrôle de qualité a extrait de sa base de données, un échantillon de 40 semaines où il a noté X, le nombre d'accidents de travail enregistrés par semaine. Il a obtenu les résultats suivants :

2 0 4 2 2 1 3 2 0 5 4 3 2 4 5 6 6 4 2 0 3 4 4 2 6 2 4 3 0 4 3 4 3 3 5 5 4 2 2 1

1. Donner le tableau statistique
2. Déterminer les paramètres de position
3. Déterminer les Paramètres de dispersion
4. Tracer le diagramme en bâtons et la courbe cumulative

Solution 4.1.2

1. Le tableau statistique :

en langage R

a=c(2,0,4,2,2,1,3,2,0,5,4,3,2,4,5,6,6,4,2,0,3,4,4,2,6,2,4,3,0,4,3,4,3,3,5,5,4,2,2,1)

z=c(table(a))

data.frame(Eff=z,ECC=cumsum(z),Frq=z/sum(z),FCC=cumsum(z/sum(z)))

```
> a=c(2,0,4,2,2,1,3,2,0,5,4,3,2,4,5,6,6,4,2,0,3,4,4,2,6,2,4,3,0,4,3,4,3,3,5,5,4,2,2,1)
> z=c(table(a))
> data.frame(Eff=z,ECC=cumsum(z),Frq=z/sum(z),FCC=cumsum(z/sum(z)))
  Eff ECC   Frq   FCC
0   4   4 0.100 0.100
1   2   6 0.050 0.150
2  10  16 0.250 0.400
3   7  23 0.175 0.575
4  10  33 0.250 0.825
5   4  37 0.100 0.925
6   3  40 0.075 1.000
>
```

FIG. 4.4 – Les résultats en R

2 Les paramètres de position :

En langage R

> a=c(2,0,4,2,2,1,3,2,0,5,4,3,2,4,5,6,6,4,2,0,3,4,4,2,6,2,4,3,0,4,3,4,3,3,5,5,4,2,2,1)

> z=c(table(a))

> n=length(a)

> xb=mean(a)

> xb

[1] 3.025

> median(a)

[1] 3

> which(z==max(z))

2 4

3 5

3 Les Paramètres de dispersion :

En langage R

> a=c(2,0,4,2,2,1,3,2,0,5,4,3,2,4,5,6,6,4,2,0,3,4,4,2,6,2,4,3,0,4,3,4,3,3,5,5,4,2,2,1)

> z=c(table(a))

> E=max(a)-min(a)#L'étendue

> E

[1] 6

> s=var(a)# la variance

> s

[1] 2.742949

> c=sd(a)# L'écart-type

> c

[1] 1.656185

> q=quantile(a)

> q

0% 25% 50% 75% 100%

0 2 3 4 6

> IQR(a)# Intervalle interquartille

[1] 2

> Coff=sd(a)/mean(a)# Le cofficient de variation

> Coff

[1] 0.5474992

> emoy=mean(abs(a-mean(z)))# L'écart absolu moyen

> emoy

[1] 2.732143

> emed=mean(abs(a-median(z)))# L'écart mediane

> emed

[1] 1.475

4 Représentations graphiques :

a=c(2,0,4,2,2,1,3,2,0,5,4,3,2,4,5,6,6,4,2,0,3,4,4,2,6,2,4,3,0,4,3,4,3,3,5,5,4,2,2,1)

z=c(table(a))

n=length(a)

plot(z,typ="h",xlab ="le nombre d'accidents",ylab ="effectif",main =" ",frame=0,lwd=3)# Le diagramme en bâtons

plot(ecdf(z),xlab ="le nombre d'accidents",ylab ="effectif",main =" ",frame=0)# La courbe cumulative

Les résultats en R

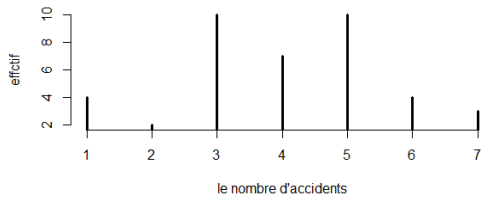


FIG. 4.5 – Le diagramme en bâtons

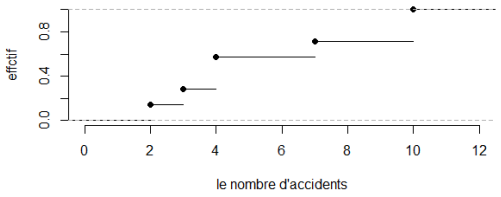


FIG. 4.6 – La courbe cumulative

4.1.3 Cas des variables quantitatives continue

Exemple 4.1.3

On prélève 20 poulets dans un élevage et on mesure le taux de dioxine ($\mu g/l$) contenu dans leur viande afin d'estimer le taux moyen pour tout l'élevage.

Les résultats sont donnés dans le tableau ci-dessous

Poulets	1	2	3	4	5	6	7	8	9	10
Taux	0.34	0.23	0.11	0.42	0.22	0.33	0.16	0.12	0.11	0.21
Poulets	11	12	13	14	15	16	17	18	19	20
Taux	0.14	0.44	0.27	0.36	0.43	0.38	0.34	0.22	0.32	0.11

1. Donner le tableau statistique
2. Déterminer les paramètres de tendance centrale
3. Déterminer les paramètres de dispersion

Solution 4.1.3

1. Le tableau statistique :

```
> T=c(0.11,0.11,0.11,0.12,0.14,0.16,0.21,0.22,0.22,0.23,0.27,0.32,0.33,0.
34,0.34,0.36,0.38,0.42,0.43,0.44)
> a=cut(T,c(0.11,0.18,0.25,0.32,0.39,0.46) ,right = FALSE)
> s=table(a)
> y=c(s)
> data.frame(Eff=y,ECC=cumsum(y),Frq=y/sum(y),FCC=cumsum(y/sum(y)))
          Eff ECC  Frq  FCC
[0.11,0.18)   6   6 0.30 0.30
[0.18,0.25)   4  10 0.20 0.50
[0.25,0.32)   1  11 0.05 0.55
[0.32,0.39)   6  17 0.30 0.85
[0.39,0.46)   3  20 0.15 1.00
```

2 Les paramètres de tendance central :

En langage R

> mean(T)

[1] 0.263

> median(T)

[1] 0.25

La classe modale 1 est $[0,11 - 0,18[$ $Mo_1 = 0,162$

La classe modale 2 est $[0,32 - 0,39[$ $Mo_1 = 0,363$

3 Les paramètres de dispersion

En langage R

> quantile(T)

0% 25% 50% 75% 100%

0.110 0.155 0.250 0.345 0.440

> var(T)

[1] 0.01319053

> sd(T)

[1] 0.11485

> cv=sd(T)/mean(T)

> cv

[1] 0.4366921

> IQR(T)

[1] 0.19

4.2 Statistique descriptive bivariée

Exemple 4.2.1

Une expérience a été faite sur 10 grenouilles pour voir s'il existe une relation entre la teneur mélanine de la peau (X) et le poids en grammes (Y)

i	1	2	3	4	5	6	7	8	9	10
X_i	0.11	0.15	0.32	0.68	0.64	0.29	0.45	0.51	0.14	0.71
Y_i	11	19	20	18	17	22	25	24	21	23

1. Déterminer la description numérique
2. Calcule la droite d'ajustement
3. Représentation graphique du nuage de points de la relation

Solution 4.2.1

1. Description numérique

En langage R

> #**Lecteur des données**

> X=c(0.11,0.15,0.32,0.68,0.64,0.29,0.45,0.51,0.14,0.71)

> Y=c(11,19,20,18,17,22,25,24,21,23)

> data.frame(X ,Y)

  X Y

1 0.11 11

2 0.15 19

3 0.32 20

4 0.68 18

5 0.64 17

6 0.29 22

7 0.45 25

8 0.51 24

9 0.14 21

10 0.71 23

> N1=length(X)

> N1

[1] 10

> N2=length(Y)

> N2

[1] 10

> #**Description numérique**

> mean(X)# *moyenne marginale de X*

[1] 0.4

> mean(Y)# *moyenne marginale deY*

[1] 20

> v1=sum((X-mean(X))^2)/N# *Variance marginale de X*

> v1

[1] 0.011985

> v2=sum((Y-mean(Y))^2)/N# *Variance marginale de Y*

> v2

[1] 3.75

> s1=sqrt(v1)#L'cart-type marginale de v1

> s1

[1] 0.109476

> s2=sqrt(v2)#L'cart-type marginale de v2

> s2

[1] 1.936492

> c=sum((X-mean(X))*(Y-mean(Y)))/N # Covariance

> c

[1] 0.068

> corr=c/(s1*s2)# Cofficient de corrélation

> corr

[1] 0.3207556

Remarque 4.2.1 *Nous remarquons que : corr* > 0 *alors il existe une relation linéaire entre* X *et* Y

2. La droite d'ajustement :$Y = bX + a$

En langage R

> X=c(0.11,0.15,0.32,0.68,0.64,0.29,0.45,0.51,0.14,0.71)

> Y=c(11,19,20,18,17,22,25,24,21,23)

> b=c/v1

> b

[1] 5.673759

> a=mean(Y)-b*mean(X)

> a

[1] 17.7305

Alors Y=5.673759X+17.7305

3. Représentation graphique du nuage de points de la relation

En langage R

> X=c(0.11,0.15,0.32,0.68,0.64,0.29,0.45,0.51,0.14,0.71)

> Y=c(11,19,20,18,17,22,25,24,21,23)

> plot(X, Y)

> x=0.11 :0.2 :0.8

> d=lm(Y~X)

> abline(d,col="red")

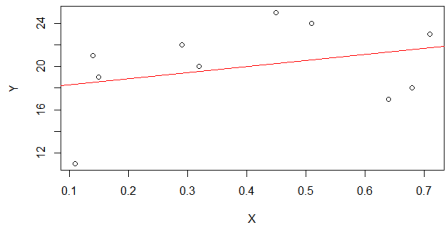


FIG. 4.7 – Le droite de régression

Conclusion

Nous avons atteint la fin de la thèse scientifique relative aux statistiques descriptives, considérées comme un outil puissant et nécessaire pour analyser et comprendre les données. J'ai essayé de développer ce sujet autant que possible, car j'ai utilisé de nombreux exemples simples et clairs pour transmettre des informations. Certains graphiques d'analyse spéciaux et des techniques modernes telles que la logique R peuvent également être utilisés pour résoudre le problème dans les plus brefs délais. Mon objectif principal dans cette thèse était de déterminer l'importance des statistiques descriptives dans notre vie quotidienne et scientifique. On peut dire que les statistiques descriptives nous aident. La description des statistiques facilite de nombreuses choses et permet de prendre des décisions fondées sur des preuves. Elle nous aide à décrire les phénomènes et à révéler des informations cachées de manière systématique et précise. Enfin de compte, j'espère à Dieu que ma thèse sera bénéfique à la prochaine génération, et je conclus ma recherche en disant au nom de Dieu, le Compatissant, le Miséricordieux "Allah élèvera en degrés ceux d'entre vous qui auront cru et ceux qui auront reçu le savoir ". AlMujadalah-11.

Al-Shafei, que Dieu ait pitié de lui, a dit : "La connaissance n'est pas préservée, mais la connaissance est bénéfique."

Bibliographie

[1] Amour, M. (2020-2021). Intitulé du module Statistique descriptive.

[2] Amrani, S. (2020-2021). statistique descriptive

[3] Ben Saïd, F. (2012-2013). Résumé de cours de Statistiques descriptives.

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

[5] Ben ameur, S. (2021-2022). Analyse de données.

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

[7] Djoudad, K. (2017).La statistique descriptive univariée applique àla biologie

[8] Daniel Fredon. Mini Manuel de Probabilités et statistique : Cours et exercices corrigés. Dunod, 2014.

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

[10] Mzali, H. (2013). Statistique et calcul de propabilite.

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

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

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

[14] Yousfi, S. (2023).Cours de Statistique Descriptive.

[15] http ://pf-mh.uvt.rnu.tn/32/1/SN1011.pdf

[16] http ://www.itse.be/statistique2010/co/Module_statistique_FSP.html.

[17] https ://www.unilim.fr/pages_perso/francisco.silva/cursoestadistica_16.pdf.

Annexe A : Logiciel R

Annexe B : Abréviations et Notations

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

Symbole		Signification
$\mathcal{P}$	:	Population l'ensemble sur lequel porte notre étude statistique.
u	:	Individu ou Unité statistique
X	:	Caractère ou variable statistique
VS	:	Variable statistique
N	:	Effectif total
$Card(\mathcal{P})$	:	Le cardinal : nombre d'éléments de l'ensemble $\mathcal{P}$.
$\sum_{i=1}^{p}$	:	La somme pour i variant de 1 à p
ECC	:	L'effectif cumulé croissants
ECD	:	L'effectif cumulé décroissant
FCC	:	Fréquences cumulées croissants
FCD	:	Fréquences cumulées décroissantes
$F(x)$	:	fonction de répartition
n_i	:	Effectif observé dans la classe i
f_i	:	Fréquence observé dans la classe i
c_i	:	Le centre d'une classe
MOC	:	Moindres carrés ordinaire
d_i	:	La densité d'effectif
δ_i	:	La densité de fréquence
IQ	:	Les intervalles interquantiles
$var(.)$	:	Variance mathématique
$\sigma_{(x)}$	:	L'cart-type de
$CV(x)$	:	Coefficient de variation de x
q_α	:	Les quantiles d'ordre (α)
E	:	l'étendue
Med	:	la médiane
$Cov(X,Y)$	:	Covariance mathématique du couple (X,Y).
M_o	:	Le mode
a_i	:	L'amplitude d'une classe
b_i , b_{i+1}	:	Les bornes d'une classe

الملخص

نستخلص من هده الأطروحة آن الإحصاء الوصفي يقوم بدراسة السلاسل الإحصائية التي تتألف من متغير واحد أو متغيرين أو أكثر .حيث يستخدم طرق لجمع المعلمات وتحليلها من خلال ربط المعلمات بيه. هذه الطرق ساعدنا في تجميع البيانات وتنظيمها وتلخيصها وعرضها بطريقة مبسطة و واضحة في صورة جداول وأشكال بيانية وحساب المقاييس الإحصائية المختلفة لوصف متغير (أو أكثر)في مجتمع ما.

Résumé

Nous déduisons de cette thèse que la statistique descriptive étudie les séries statistiques composées d'une seule variable ou de plusieurs. Elle utilise des méthodes pour collecter et analyser les paramètres en les reliant entre eux. Ces méthodes nous ont aidés à recoller, organiser, résumer et présenter les données de manière simple et claire sous forme de tableaux et de graphiques, ainsi qu'à calculer différentes mesures statistiques pour décrire une ou plusieurs variables dans une population donnée.

Summary

From this thesis, we can deduce that descriptive statistics studies statistics composed of a single variable or multiple variables. It uses methods to collect and analyze parameters by linking them together. These methods have helped us to collate, organize, summarize, and present data in a simple and clear manner in the form of tables and graphs. Additionally, they are used to calculate various statistical measures to describe one or more variables in a given population.