REPUBLIQUE ALGERIENNE DÉMOCRATIQUE & POPULAIRE.

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DJILALI BOUNAAMA - KHEMIS MELIANA

FACULTÉ DES SCIENCES & TECHNOLOGIE

Département de Mathématique & Informatique

POLYCOPIÉ DE COURS INTITULÉ :

ANALYSE DE DONNÉES
COURS ET EXCERCICES

POLYCOPIÉ DESTINÉ AUX ÉTUDIANTS 1ere ANNEE MASTER INFORMATIQU

Présenté par : Mme. BERKANE Kheira

Maître Assistant B

2021/2022

Table des matières

Introduction

Ce polycopié s'adresse aux étudiants de master Informatique et à tous ceux désirant s'initier et faire un tour sur les principaux méthodes d'analyse des données .

Il est conseillé aux lecteurs d'avoir de bonnes connaissances d'Algèbre linéaire (Calcul matricielle, diagonalisaton, normes matricielles) et quelques notions d'Analyse mathématiques (espace métrique, produit scalaire, recherche d'extrema ....).

L'expression " Analyse des Données" recouvre les techniques utilisées pour décrire les grands tableaux. Ces techniques regroupent un certain nombre d'outils statistiques permettant de construire des supports et/ou des résumés de l'information afin de faciliter l'interprétation .

Parmi ces outils, ce trouvent les méthodes dites factorielles qui fournissent des représentations graphiques sous la forme de nuage de points provenant de projections sur des plans choisis. Ces méthodes ont le gros avantage de traiter à la fois les individus et les variables. et les méthodes de classification.

Les méthodes factorielles que nous aborderons dans ce document, sont l'Analyse en Composante Principale, ou ACP ; l'Analyse Factorielle des Correspondances ou AFC ; ainsi que l'Analyse des Correspondances Multiples .

Nous commençons par aborder l'ACP dans le premier chapitre. Les deux suivants chapitres sur l'Analyse Factorielle simple (AFC) et Multiple (AFCM) qui ne peuvent être abordés sans avoir assimilé correctement le chapitre sur l'ACP, ensuite on termine par quelques méthodes/technique de classification dans le dernier chapitre .

Tout au long du document des exemples et des exercices de cours sont proposés dont le corrigé est parfois fourni avec l'excercice.

Les TPs de ce module sont réalisés soit avec le logiciel Matlab où le logiciel libre R. ils sont des logiciels de statistique et d'analyse des Données. Ils possèdent de nombreuses

fonctionnalités pour la modélisation et la visualisation de données.

Rappel sur la statistique descriptive

1. Vocabulaire :

- **Population** : c'est l'ensemble étudié.
- **Individu** : c'est un élément de la population.
- **Effectif total** : c'est le nombre total d'individus.
- **Caractère** : c'est la propriété étudiée.
  On distingue les **caractères discrets** qui ne peuvent prendre qu'un nombre fini de valeurs (notes à un devoir...) et les **caractères continus** dont on regroupe les valeurs par intervalles (taille, durée d'écoute...).

2. Séries statistiques associées à un caractère discret

1) Classement des données

Définition
On appelle **série statistique** la donnée simultanée (dans un tableau) des valeurs du caractère étudié (noté x_i), rangées dans l'ordre croissant, et des effectifs (notés n_i) de ces valeurs.

► **Remarque :** A la place des effectifs (n_i), on peut aussi utiliser les fréquences $f_i = \dfrac{n_i}{N}$ (où N représente l'effectif total) ou les fréquences en pourcentages $f_i = \dfrac{n_i}{N} \times 100$.

► *Exemple :* Les notes sur 20 obtenues lors d'un devoir de mathématiques dans une classe de seconde sont les suivantes :
10, 8, 11, 9, 12, 10, 8, 10, 7, 9, 10, 11, 12, 10, 8, 9, 10, 9, 10, 11.
• La population étudiée est la classe et les individus sont les élèves. L'effectif total est égal à 20 et la note obtenue au devoir est le caractère discret que l'on étudie.
• La série statistique définie par les effectifs est la suivante :

Valeurs du caractère (notes) x_i	7	8	9	10	11	12
Effectifs (nb d'élèves ayant la note) n_i	1	3	4	7	3	2

• La série statistique définie par les fréquences en pourcentage est la suivante :

Valeurs du caractère (notes) x_i	7	8	9	10	11	12
Fréquences en % $f_i = \frac{n_i}{20} \times 100$	5 %	15 %	20 %	35 %	15 %	10 %

2) Effectifs cumulés

Définition
L'effectif cumulé croissant d'une valeur x est la somme des effectifs des valeurs y tels que $y \leqslant x$.
L'effectif cumulé décroissant d'une valeur x est la somme des effectifs des valeurs y tels que $y > x$.

► *Avec l'exemple des notes, on a :*

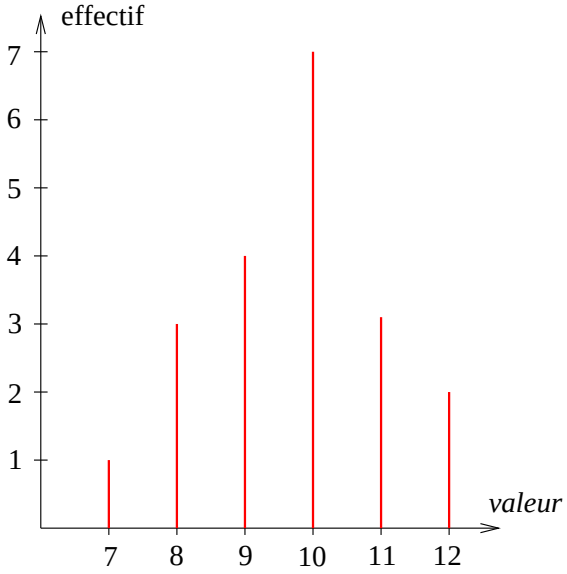
Valeurs x_i	7	8	9	10	11	12
Effectif cumulé croissant	1	4*	8	15	18	20
Effectif cumulé décroissant	19	16**	12	5	2	0

* : nombre d'élèves ayant eu une note $\leqslant 8$; ** : nombre d'élèves ayant eu une note > 8

3) Représentation graphique

Pour les caractères quantitatifs discrets, on utilise le **diagramme en bâton** :
Dans un repère orthogonal, pour chaque valeur de la série statistique on trace un trait vertical dont la hauteur est proportionnelle à l'effectif (dans l'unité choisie).

► *Avec l'exemple des notes :*



4) Paramètres de position

a) Moyenne

Définition

On appelle **moyenne** d'une série statistique d'effectif total N, le réel $\overline{x} = \dfrac{n_1x_1 + n_2x_2 + \cdots + n_kx_k}{N}$.

(k représente le nombre de valeurs prises par le caractère)

► *Avec l'exemple des notes, on a :*

Valeurs du caractère x_i	7	8	9	10	11	12
Effectifs n_i	1	3	4	7	3	2

$$\overline{x} = \frac{1\times 7 + 3\times 8 + 4\times 9 + 7\times 10 + 3\times 11 + 2\times 12}{20} = 9,7$$

► **Remarques :**

• En utilisant les fréquences, on a : $\overline{x} = f_1x_1 + f_2x_2 + \cdots + f_kx_k$.

• Avec les fréquences en pourcentages, on a : $\overline{x} = \dfrac{f_1x_1 + f_2x_2 + \cdots + f_kx_k}{100}$.

Propriété

- Si on ajoute à toutes les valeurs d'une série statistique le même nombre b, on augmente la moyenne de cette série par b.
- Si les valeurs d'une série statistique sont multipliées ou divisées par un même nombre a, la moyenne de cette série est aussi multipliée ou divisée par a.

Propriété

Si une population d'effectif N est composée d'une partie d'effectif N_1 et de moyenne $\overline{x_1}$ et d'une autre partie d'effectif N_2 et de moyenne $\overline{x_2}$, alors la moyenne $\overline{x}$ de la population totale est telle que :

$$\overline{x} = \frac{N_1\overline{x_1} + N_2\overline{x_2}}{N}$$

► *Exemple :* Si dans une classe, les 15 garçons d'une classe mesurent en moyenne 182 cm et si les 20 filles mesurent en moyenne 168 cm, alors la taille moyenne d'un élève de cette classe est égale à $\dfrac{15\times 182 + 20\times 168}{15+20} = 174$ cm.

b) Médiane

Définition

L'idée générale est que la médiane est une valeur du caractère qui partage la population en deux parties de même effectif.

De façon plus précise, on appelle **médiane** d'une série statistique discrète toute valeur M du caractère telle qu'au moins 50% des individus aient une valeur du caractère inférieure ou égale à M et au moins 50% des individus aient une valeur du caractère supérieure ou égale à M.

Recherche pratique de la médiane :

On range les valeurs du caractère une par une dans l'ordre croissant (chaque valeur du caractère doit apparaître un nombre de fois égal à l'effectif correspondant).

Si l'effectif total est impair, la médiane M est la valeur du caractère située au milieu.

Si l'effectif total est pair, la médiane M est la demi-somme des 2 valeurs situées au milieu.

► *Exemple 1 :*

On considère la série statistique suivante :

Valeurs du caractère x_i	7	8	9	10	11	14	16
Effectifs n_i	2	1	1	1	2	1	2

• Liste des valeurs du caractère :

7 ; 7 ; 8 ; 9 ; $\underbrace{10 ; 11}$; 11 ; 14 ; 16 ; 16

• L'effectif total est pair : la médiane M est la demi-somme des 2 valeurs situées au milieu. D'où, $M = \dfrac{10+11}{2} = 10,5$.

► *Exemple 2 :*

On considère la série statistique suivante :

Valeurs du caractère x_i	6	8	9	12	13	17
Effectifs n_i	3	1	2	1	3	3

• Liste des valeurs du caractère :

6 ; 6 ; 6 ; 8 ; 9 ; 9 ; $\underbrace{12}$; 13 ; 13 ; 13 ; 17 ; 17 ; 17

• L'effectif total est impair : la médiane M est la valeur située au milieu. D'où, $M = 12$.

5) Paramètres de dispersion

Ces paramètres permettent de mesurer la façon dont les valeurs du caractère sont réparties autour de la moyenne et de la médiane.

a) Paramètre de dispersion associé à la moyenne

Définition

- On appelle **variance** d'une série statistique d'effectif total N et de moyenne $\overline{x}$, le réel $V = \dfrac{n_1(x_1-\overline{x})^2 + n_2(x_2-\overline{x})^2 + \cdots + n_k(x_k-\overline{x})^2}{N}$ (moyenne des carrés des écarts à la moyenne)
- **l'écart-type** de la série est défini alors par : $\sigma = \sqrt{\quad}$

–

Valeurs du caractère x_i	7	8	9	10	11	12
Effectifs n_i	1	3	4	7	3	2

$$V = \frac{1\times(7-9,7)^2 + 3\times(8-9,7)^2 + 4\times(9-9,7)^2 + 7\times(10-9,7)^2 + 3\times(11-9,7)^2 + 2\times(12-9,7)^2}{20} = 1,71$$

$\sigma = \sqrt{1,71} \approx 1,31$

Propriété

- Si on ajoute à toutes les valeurs d'une série statistique le même nombre b, l'écart-type reste inchangé.
- Si les valeurs d'une série statistique sont multipliées ou divisées par un même nombre strictement positif a, l'écart-type est multiplié ou divisé par a.
- Si les valeurs d'une série statistique sont multipliées ou divisées par un même nombre strictement négatif a, l'écart-type est multiplié ou divisé par $-a$.

b) Paramètre de dispersion associé à la médiane

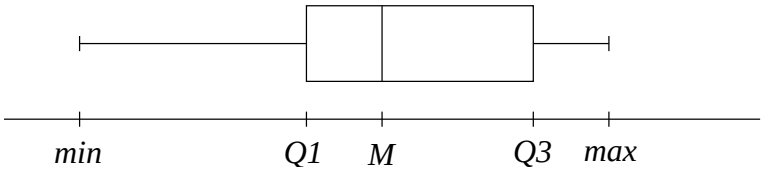
Définition

L'idée générale est de partager la population en quatre parties de même effectif.
Etant donné une série statistique de médiane M dont la liste des valeurs est rangée dans l'ordre croissant *(il s'agit de la même liste que celle qu'on utilise pour déterminer la médiane)*.
En coupant la liste en deux sous-séries de même effectif *(Attention : quand l'effectif total est impair, la médiane ne doit pas être incluse dans les sous-séries)* :

- On appelle **premier quartile** le réel noté Q_1 égal à la médiane de la sous-série inférieure.
- On appelle **troisième quartile** le réel noté Q_3 égal à la médiane de la sous-série supérieure.
- **L'écart interquartile** est égal à $Q_3 - Q_1$.
- $]Q_1 ; Q_3[$ est appelé **intervalle interquartile**.

Définition

Le **diagramme en boîtes** d'une série statistique se construit alors de la façon suivante :
(les valeurs du caractère sont en abscisse - min et max représentent les valeurs minimales et maximales du caractère)



► **Interprétation :**

- 25% de la population admet une valeur du caractère entre *min* et Q_1
- 25% de la population admet une valeur du caractère entre Q_1 et M
- 25% de la population admet une valeur du caractère entre M et Q_3
- 25% de la population admet une valeur du caractère entre Q_3 et *max*

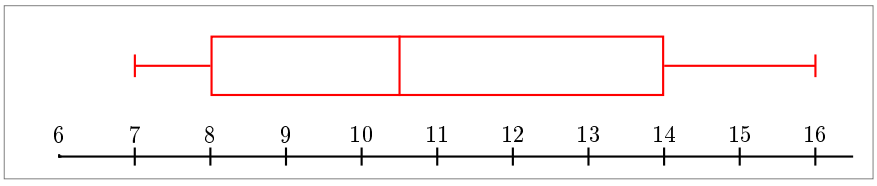
► *Exemple 1 :*
On reprend la série statistique suivante :

Valeurs du caractère x_i	7	8	9	10	11	14	16
Effectifs n_i	2	1	1	1	2	1	2

• Liste des valeurs du caractère :

$$\underbrace{7\,;\,7\,;\,\overbrace{8}\,;\,9\,;\,10}_{\text{sous-série inférieure}}\,;\,\underbrace{11\,;\,11\,;\,\overbrace{14}\,;\,16\,;\,16}_{\text{sous-série supérieure}}$$

• L'effectif de chaque sous-série est impair : $Q_1 = 8$ (valeur située au milieu de la sous-série inférieure) et $Q_3 = 14$ (valeur située au milieu de la sous-série supérieure).
• Le diagramme en boîtes de la série est le suivant :



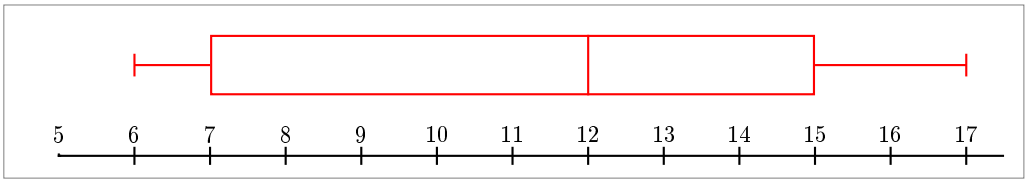
► *Exemple 2 :*
On reprend la série statistique suivante :

Valeurs du caractère x_i	6	8	9	12	13	17
Effectifs n_i	3	1	2	1	3	3

• Liste des valeurs du caractère :

$\underbrace{6\,;\,6\,;\,\overbrace{6\,;\,8}\,;\,9\,;\,9}_{\text{sous-série inférieure}}\,;\,12\,;\,\underbrace{13\,;\,13\,;\,\overbrace{13\,;\,17}\,;\,17\,;\,17}_{\text{sous-série supérieure}}$

• L'effectif de chaque sous-série est pair : $Q_1 = 7$ (demi-somme des deux valeurs situées au milieu de la sous-série inférieure) et $Q_3 = 15$ (demi-somme des deux valeurs situées au milieu de la sous-série supérieure).
• Le diagramme en boîtes de la série est le suivant :



3. Séries statistiques associées à un caractère continu

1) Classement des données

La seule différence par rapport aux caractères discrets, c'est que les valeurs du caractère sont regroupées dans des intervalles (appelés classes du caractère).
► *Exemple :* Temps passé devant la télévision par 34 élèves pendant une certaine journée.

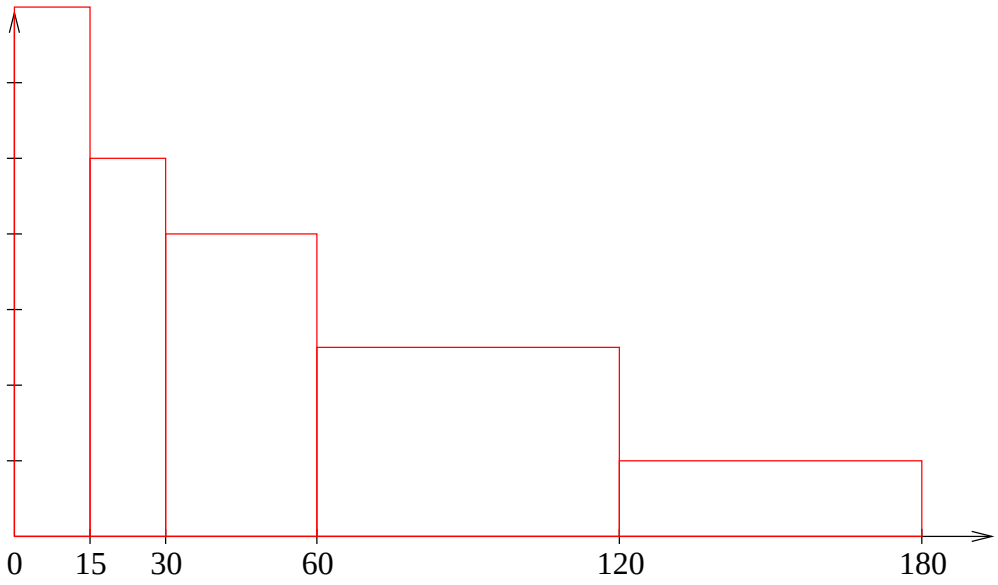
temps en minutes	$[0,15[$	$[15,30[$	$[30,60[$	$[60,120[$	$[120,180[$
nombre d'élèves	7	5	8	10	4

2) Représentation graphique

Pour la représentation graphique d'un caractère continu, on utilise généralement un **histogramme** : dans un repère orthogonal on porte en abscisse les valeurs des bornes des intervalles (selon l'unité choisie), puis pour chaque intervalle on trace un rectangle dont **l'aire est proportionnelle à l'effectif** (selon l'unité choisie).

► **Remarque :** En pratique, il est conseillé de commencer par construire un tableau donnant la largeur et l'aire de chaque rectangle (selon les unités choisies). On peut alors facilement en déduire la hauteur de chaque rectangle ce qui facilite la construction graphique de l'histogramme.
► *Pour l'exemple proposé ci-dessus :* (unités : en abscisse 1 cm représente 15 min et 1 cm^2 représente 1 élève)

temps en minutes	$[0,15[$	$[15,30[$	$[30,60[$	$[60,120[$	$[120,180[$
aire du rectangle en cm^2	7	5	8	10	4
largeur du rectangle en cm	1	1	2	4	4
hauteur du rectangle en $cm = \dfrac{\text{aire}}{\text{largeur}}$	7	5	4	2,5	1



3) Calcul des paramètres de position et de dispersion

Pour calculer les différents paramètres d'une série statistique associé à un caractère continu, on prend comme valeur du caractère **le milieu de chaque classe**.

► *Pour l'exemple, la série devient :*

valeur (milieu de chaque intervalle) x_i	7,5	22,5	45	90	150
effectif n_i	7	5	8	10	4

On en déduit que :

$$\overline{x} = \frac{7 \times 7,5 + 5 \times 22,5 + 8 \times 45 + 10 \times 90 + 4 \times 150}{34} \approx 60$$