Université Mohammed V - Agdal
Faculté des Sciences
Département de Mathématiques
Avenue Ibn Batouta, B.P. 1014
Rabat, Maroc

Filière : SMI
Module 18

Cours Statistique Descriptive

Par

HAKAM Samir

Année : 2014 - 2015

Table des matières

Introduction

En présence d'un ensemble de données chiffrées, on a un désir spontané de simplification. Selon des critères, la statistique cherche d'une part à représenter, ordonner et classer des données ; d'autre part, à résumer la multiplicité et la complexité des notions par des caractéristiques synthétiques.

Le statisticien est ainsi conduit à collecter des données, construire des graphiques, déterminer des caractéristiques centrale et calculer des caractéristiques de dispersion.

L'organisation, la description et la présentation des données sous forme de tableaux ou de graphiques sont l'objet de la " *statistique descriptive*". L'interprétation et les conclusions que l'on peut tirer d'un ensemble de données font l'objet de la "*statistique Inférentielle*"

Chapitre 1

Distribution statistique

1.1 Généralités

1.1.1 Population

Toute étude statistique concerne un ensemble Ω appelé **population** dont les éléments sont appelés des **individus**.

Définition 1.1.1 *:* Une population c'est l'ensemble d'individus ou d'objets qui possèdent un ou plusieurs caractères spécifiques en commun.

Une population statistique est dite finie si l'on peut déterminer avec précision le nombre d'individus qui la composent sinon elle est dite infinie.

Exemple 1.1.1 *:* **i)** Dans une étude sur le sport, la population peut être l'ensemble des personnes qui pratiquent un sport.

ii) Dans une étude sur les revenus mensuels dans une entreprise, la population peut être l'ensemble des personnes qui travaillent dans cette entreprise.

1.1.2 Variables statistiques

L'étude statistique consiste en l'analyse d'une **variable** X appelé parfois caractère qui sert à décrire l'aspect d'une population objet de l'étude. On distigue deux types de variables : **qualitatives et quantitatives**.

Définition 1.1.2 *:* Une variable X est dite qualitative si les valeurs prises sont des mots ou des lettres.

Une variable X est dite quantitative si les valeurs prises sont des nombres réels.

Exemple 1.1.2 *:* On considère la population marocaine Ω.

i) La couleur des cheveux, le fait de posséder une voiture ou non définissent des variables qualitatives.

ii) La taille , le poids et l'âge sont des variables quantitatives.
On distingue deux types de variables quantitatives, **discrète** et **continue**

Définition 1.1.3 *:* Une variable quantitative X est dite discrète si les valeurs qu'elle peut prendre sont isolées les unes des autres.

Une variable quantitative X est dite continue si elle peut prendre toutes les valeurs d'un intervalle de $\mathbb{R}$ ou une réunion d'intervalles de $\mathbb{R}$ ou l'ensemble des réels $\mathbb{R}$.

Exemple 1.1.3 *:* **i)** Les performances en saut en hateurs de 100 athlètes est une variable quantitative discrète.

ii) La consommation en carburant aux 100 km d'un nouveau modèle d'une voiture est une variable quantitative continue.

1.1.3 Echantillon

Pour obtenir un renseignement exact concernant une variable X, il faut étudier tous les individus de la population. Quand cela n'est pas possible, on restreint l'étude à une partie de la population appelée **échantillon**.

Définition 1.1.4 *:* Un échantillon est une partie finie représentative de la population c'est donc un sous ensemble E de Ω.

1.2 Présentation des données statistiques

1.2.1 Effectifs - Fréquences - Fréquences cumulées

L'étude concrète d'une variable X donne N valeurs qui constituent la distribution statistique de X (aussi appelé série statistique).
Cette distribution est, en générale, présentée d'une façon groupée :

• Sous la forme $\{(x_i,\, n_i) \,/\, 1 \leq i \leq p\}$ dans le cas d'une variable qualitative ou quantitative discrète (avec $x_1 < x_2 < \cdots < x_p$ dans le cas d'une variable quantitative discrète).

• Sous la forme d'intervalles ou de classes $\{(]x_i,\, x_{i+1}],\, n_i) \,/\, 1 \leq i \leq p\}$ dans le cas d'une variable quantitative continue .

Définition 1.2.1 *:* **i) l'effectif** n_i est le nombre d'individus de la population ou de l'échantillon pour lesquels X prend la valeur x_i (dans le cas d'une variable qualitative ou quantitative discrète) ou une valeur de l'intervalle $]x_i,\, x_{i+1}]$ (dans le cas d'une variable quantitative continue).
La somme des effectifs est appelée la taille de la population ou de l'échantillon et est notée N. $N = n_1 + n_2 + \cdots + n_p$

ii) On appelle **fréquence** de la valeur x_i ou de la classe $]x_i,\, x_{i+1}]$ le nombre réel

$f_i = \dfrac{n_i}{N}$ On a évidement $\sum_{i=1}^{p} f_i = 1$

C'est la proportion de l'effectif d'une valeur de la variable par rapport à N la taille totale de la population ou de l'échantillon.

iii) On appelle **fréquence cumulée** de la valeur x_i ou de la classe $]x_i,\ x_{i+1}]$ le nombre réel

$$F(x) = \sum_{\{i/x_i \leq x\}} f_i$$

C'est la proportion des unités statistiques de la population ou de l'échantillon qui possèdent une valeur inférieure ou égale à une valeur x donnée d'une variable **quantitative**.

Exemple 1.2.1 *:* **i)**Variable qualitative : La répartition des adultes d'une résidence selon le niveau d'instruction.

Niveau d'instruction	effectifs n_i	fréquences f_i
Primaire	36	0.11
Secondaire	81	0.25
Universitaire	208	0.64
Total	$N = 325$	1

ii) Variable quantitative discrète : Les performances en saut en hauteur (en cm) de 10 athlètes sont : 191, 194, 197, 191, 200, 203, 200, 197, 203, 203.

Hauteur en cm	effectifs n_i	fréquences f_i	fréquences cumulées $F(x)$
191	2	0.2	0.2
194	1	0.1	0.3
197	2	0.2	0.5
200	2	0.2	0.7
203	3	0.3	1
Total	$N = 10$	1	

iii) Variable quantitative continue : Etude de la consommation aux 100 km de 20 voiture d'un nouveau modèle : 5.56, 5.35, 5.98, 5.77, 5.18, 5.66, 5.28, 5.11, 5.58, 5.49, 5.59, 5.33, 5.55, 5.45, 5.76, 5.23, 5.57, 5.52, 5.8, 6.0.

Consommation en litre	effectifs n_i	fréquences f_i	fréquences cumulées $F(x)$
$[5, 5.2]$	2	0.1	0.1
$]5.2, 5.4]$	4	0.2	0.3
$]5.4, 5.6]$	8	0.4	0.7
$]5.6, 5.8]$	4	0.2	0.9
$]5.8, 6]$	2	0.1	1
Total	$N = 20$	1	

1.2.2 Distribution statistique

Définition 1.2.2 *:* Une distribution statistique est une représentation des données collectées dans un tableau où figurent les valeurs que prenne la variable, les effectifs, les fréquences et les fréquences cumulées relatives à chaque valeur ou ensemble de valeurs prises par la variable.

1.3 Représentations graphiques

1.3.1 Représentations graphiques d'une distribution de variables qualitatives

1.3.1.1 Les tuyaux d'orgues

Les tuyaux d'orgues des effectifs (respectivement des fréquences) de la distribution statistique $\{(x_i, n_i) / 1 \leq i \leq p\}$ (respectivement $\{(x_i, f_i) / 1 \leq i \leq p\}$) s'obtient en traçant sur un repére orthonormé, pour tout $i = 1, \cdots, p$, un rectangle de base de centre x_i et de hauteur égale à l'effectif ou la fréquence de la valeur x_i.
Sur l'axe des abscisses on représente les modalités de la variable, alors que sur l'axe des ordonnées on représente les effectifs ou les fréquences selon que l'on désire tracer un diagramme des effectifs ou des fréquences.

Exemple 1.3.1 *:* Représentation du diagramme en tuyaux d'orgues des fréquences pour le niveau d'étude des adultes d'une résidence.

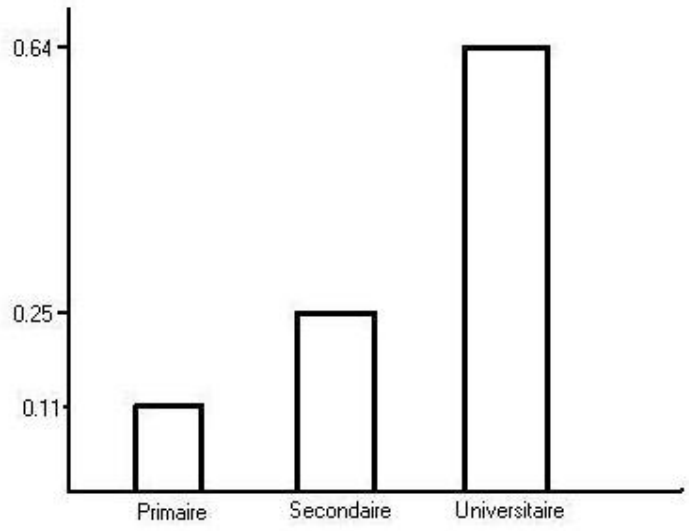


FIGURE 1.1 – Diagramme en tuyaux d'orgues

1.3.1.2 Représentation circulaire

C'est une représentation où chaque modalité est représentée par une portion du disque. Si S est l'aire du disque, l'aire d'une portion est égale à $f \times S$, où f est la fréquence de la modalité correspondante.
L'angle α de chaque portion s'obtient en multipliant la fréquence par 360°, l'angle du disque ($\alpha = f \times 360$)

Exemple 1.3.2 *:* Représentation du digramme circulaire des fréquences pour le niveau d'étude des adultes d'une résidence.

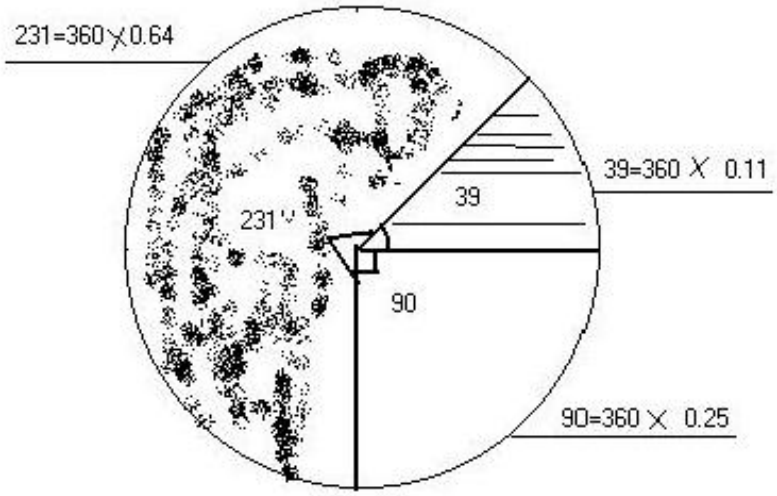


FIGURE 1.2 – Diagramme circulaire

1.3.2 Représentations graphiques d'une distribution de variables quantitatives discrètes

1.3.2.1 Diagramme en bâtons

Le diagramme en bâtons des effectifs (respectivement des fréquences) de la distribution statistique $\{(x_i, n_i) / 1 \leq i \leq p\}$ (respectivement $\{(x_i, f_i) / 1 \leq i \leq p\}$) s'obtient en traçant sur un repére orthonormé les "bâtons" A_iB_i, c'est à dire les segments joignant les point $A_i(x_i, 0)$ et $B_i(x_i, n_i)$ (respectivement $B_i(x_i, f_i)$) pour $1 \leq i \leq p$.
Sur l'axe des abscisses on représente les valeurs de la variable, alors que sur l'axe des

ordonnées on représente les effectifs ou les fréquences selon que l'on désire tracer un diagramme des effectifs ou des fréquences.

Exemple 1.3.3 *:* La distribution des performances en saut en hauteur de 100 athlètes sont représentées dans le tableau suivant :

Hauteur en cm	effectifs n_i	fréquences f_i	fréquences cumulées $F(x)$
191	6	0.06	0.06
194	17	0.17	0.23
197	41	0.41	0.64
200	27	0.27	0.91
203	9	0.09	1
Total	100	1	

Représentation du diagramme en bâtons pour la distribution des performances en saut en hauteur de 100 athlètes.

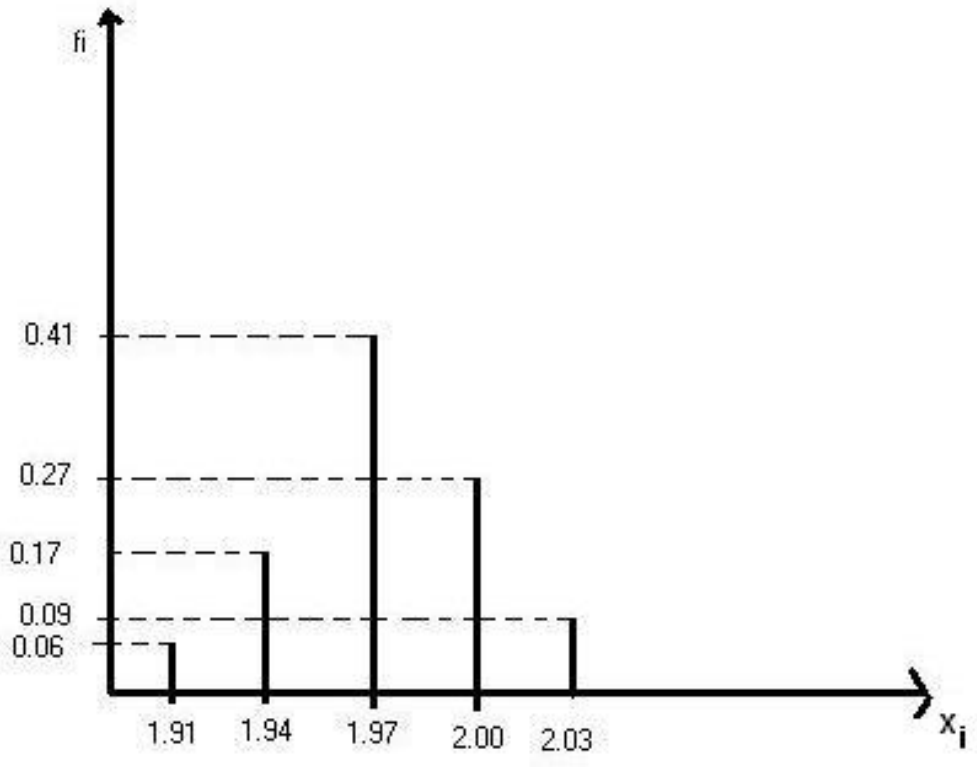


FIGURE 1.3 – Diagramme en bâtons

1.3.2.2 Polygône des fréquences

C'est une ligne brisée joignant les points de coordonnées (x_i, f_i). C'est aussi la ligne qui joint les sommets des bâtons du diagramme.

Exemple 1.3.4 *:* Représentation du polygône des fréquences pour la distribution des performances en saut en hauteur de 100 athlètes.

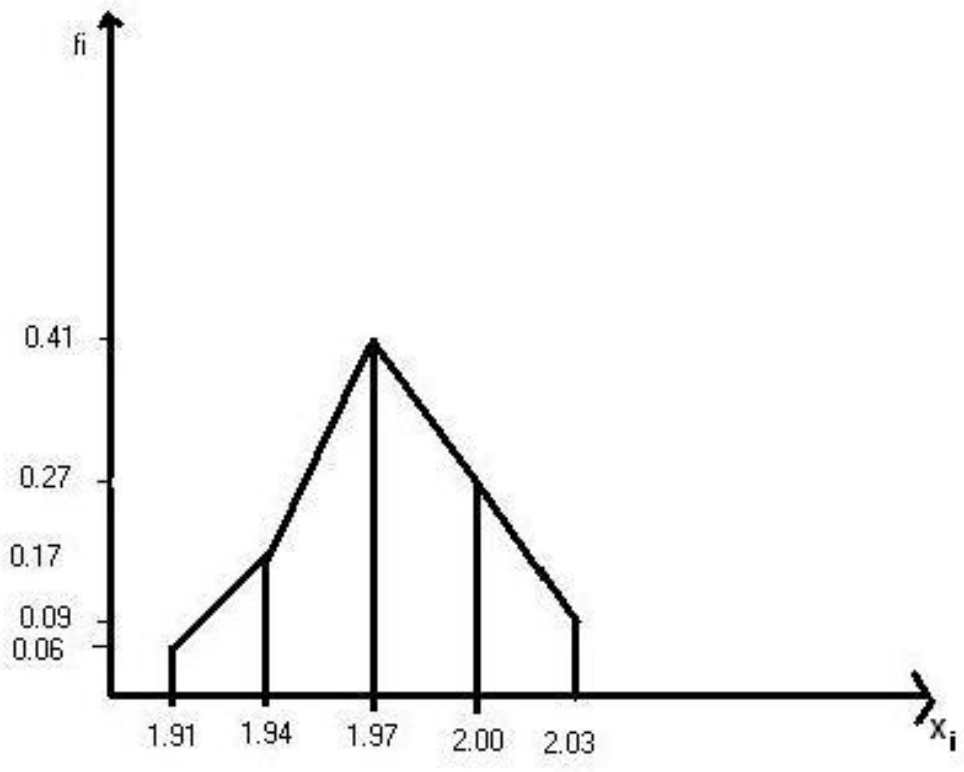


FIGURE 1.4 – Polygône des fréquences

1.3.2.3 Courbe des fréquences cumulées

C'est une courbe en escaliers qui représente la fonction :

$$F(x) = 0 \text{ si } x < x_1 \text{ et } F(x) = \sum_{j:x_j \leq x} f_j \text{ sinon}$$

Exemple 1.3.5 *:* Représentation de la courbe des fréquences cumulées pour la distribution des performances en saut en hauteur de 100 athlètes.

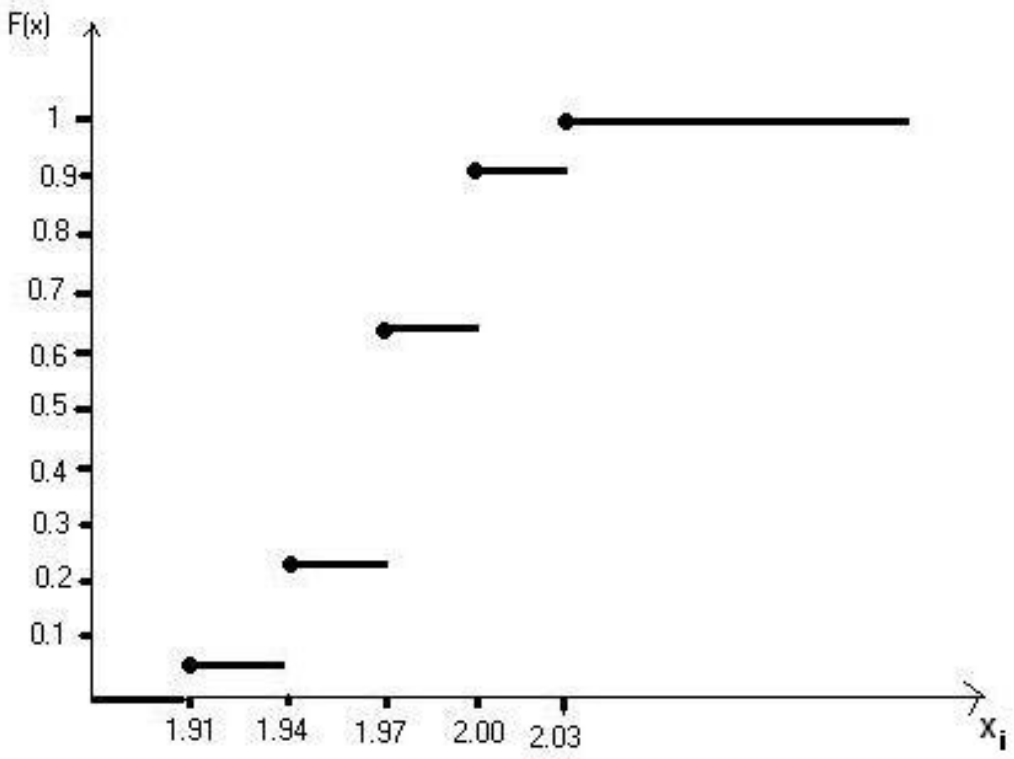


FIGURE 1.5 – Courbe des fréquences cumuleés

1.3.3 Représentations graphiques d'une distribution de variables quantitatives continues

Considérons une variable continue X dont les valeurs se situent dans un intervalle I. On divise cet intervalle en k classes disjointes $]a_i, a_{i+1}]$, $i = 1, ..., p$. On prendra toujours des classes de même amplitude ($a_{i+1} - a_i =$ constante).

Pour tout i, on note n_i le nombre de valeurs de X dans la classe $]a_i, a_{i+1}]$ qu'on appelle effectif de cette classe.

Le choix du nombre de classes est laissé au soin de l'utilisateur. Plus le nombre d'observations est grand plus le nombre de classes est élévé. On admet cependant, pour aider à la compréhention, que ce nombre devrait être entre 5 et 15. Pour dresser le tableau de distribution des effectifs et des fréquences on pourra suivre les étapes suivantes :

Etape 1 : Déterminer p le nombre de classes à considérer dans l'étude.

Etape 2 : Calculer l'étendue $e = x_{max} - x_{min}$ où x_{min} est la valeur minimale de la variable X et x_{max} est la valeur maximale de la variable X.

Etape 3 : Diviser l'étendue e par p le nombre de classes, pour avoir une idée sur la valeur de l'amplitude des classes que l'on notera a. on a

$$a = \frac{e}{p}$$

Etape 4 : On construit alors les classes

$$[x_{min}, x_{min}+a],\ ]x_{min}+a, x_{min}+2a],\ \cdots,\ ]x_{min}(p-1)\,a, x_{min}+p\,a]$$

Etape 5 : S'assurer que chaque observation appartient à une et une seule classe.

Exemple 1.3.6 *:* Etude de la consommation aux 100 km de 20 voitures d'un nouveau modèle : 6.11, 6.05, 5.98, 5.77, 5.18, 5.66, 5.28, 5.11, 5.58, 5.49, 5.62, 5.33, 5.55, 5.45, 5.76, 5.23, 5.57, 5.52, 5.8, 6.0.

Supposons que l'on décide que le nombre de classes est $p = 5$.

Nous avons $x_{min} = 5.11$ et $x_{max} = 6.11$. D'ou

$$e = 6.11 - 5.11 = 1 \text{ et } a = \frac{e}{p} = \frac{1}{5} = 0.2$$

Consommation en litre	effectifs n_i	fréquences f_i	fréquences cumulées $F(x)$
$[5.11, 5.31]$	4	0.2	0.2
$]5.31, 5.51]$	3	0.15	0.35
$]5.51, 5.71]$	6	0.3	0.65
$]5.71, 5.91]$	3	0.15	0.8
$]5.91, 6.11]$	4	0.2	1
Total	20	1	

1.3.3.1 Histogramme

L'histogramme des effectifs (respectivement des fréquences) de la distribution statistique $\{(]a_i,\, a_{i+1}],\, n_i) \,/\, 1 \leq i \leq p\}$ (respectivement $\{(]a_i,\, a_{i+1}],\, f_i) \,/\, 1 \leq i \leq p\}$) s'obtient en traçant sur un repère orthonormé, pour tout $i = 1, \cdots, p$, un rectangle de base la longueur du segment $]a_i,\, a_{i+1}]$ et de hauteur égale à l'effectif ou la fréquence de cette classe. Sur l'axe des abscisses on représente les bornes des classes $]a_i,\, a_{i+1}]$ de la variable c'est à dire les points $a_1,\, a_2, \cdots, a_p,\, a_{p+1}$, alors que sur l'axe des ordonnées on représente les effectifs ou les fréquences selon que l'on désire tracer un histogramme des effectifs ou des fréquences.

Exemple 1.3.7 *:* Représentation de l'histogramme des fréquences de la distribution de l'exemple 1.3.6.

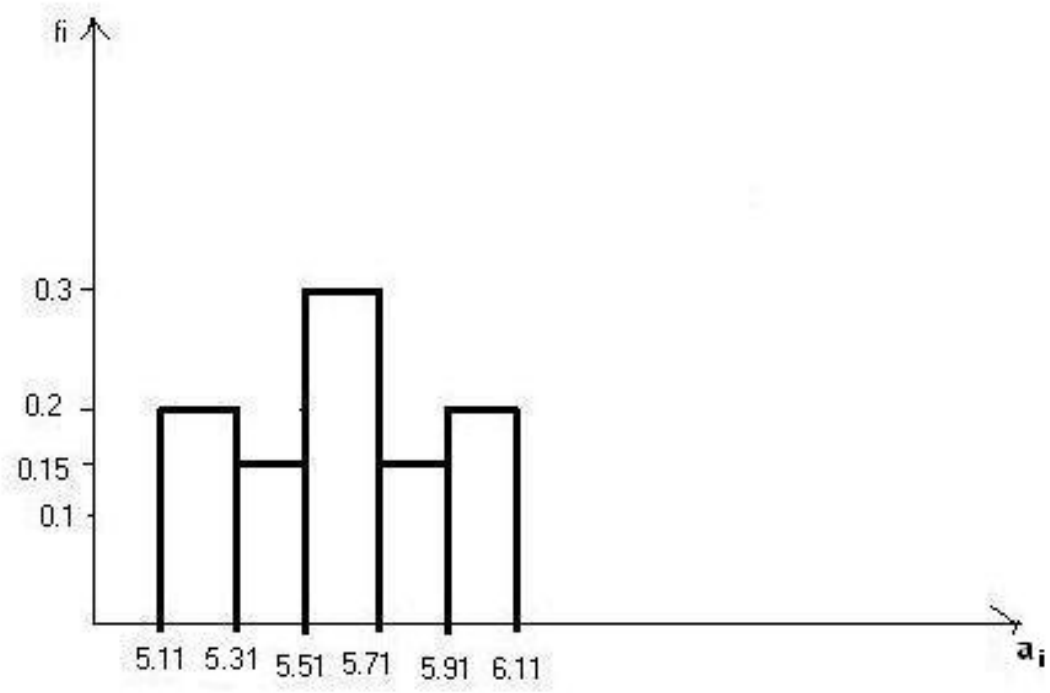


FIGURE 1.6 – Histogramme

1.3.3.2 Polygône des fréquences

Le polygône des fréquences de la distribution $\{(]a_i,\, a_{i+1}],\, f_i) \,/\, 1 \leq i \leq p\}$ est la ligne brisée joignant les points de coordonnées (c_i, f_i) où $c_i = \dfrac{a_i + a_{i+1}}{2}$, $i = 1, \cdots, p$.
Lorsque la borne inférieure de la première classe est observée c'est à dire l'intervalle est fermé en a_1 (comme c'est le cas dans l'exemple 1.3.6), on complète la courbe en joignant les points $(c_0,\, 0)$ et (c_1, f_1) où $c_0 = a_1 - \dfrac{a}{2}$.
Lorsque la borne inférieure de la première classe n'est pas observée c'est à dire l'intervalle est ouvert en a_1, on complète la courbe en joignant les points $(a_1,\, 0)$ et (c_1, f_1).
Lorsque la borne supérieure de la dernière classe est observée c'est à dire l'intervalle est fermé en a_{p+1} (comme c'est le cas dans l'exemple 1.3.6), on complète la courbe en joignant les points $(c_p,\, f_p)$ et $(c_{p+1}, 0)$ où $c_{p+1} = a_{p+1} + \dfrac{a}{2}$.
Lorsque la borne supérieure de la dernière classe n'est pas observée c'est à dire l'intervalle est ouvert en a_{p+1}, on complète la courbe en joignant les points $(c_p,\, f_p)$ et $(a_{p+1}, 0)$.

Exemple 1.3.8 *:* Représentation du polygône des fréquences de la distribution de l'exemple 1.3.6.

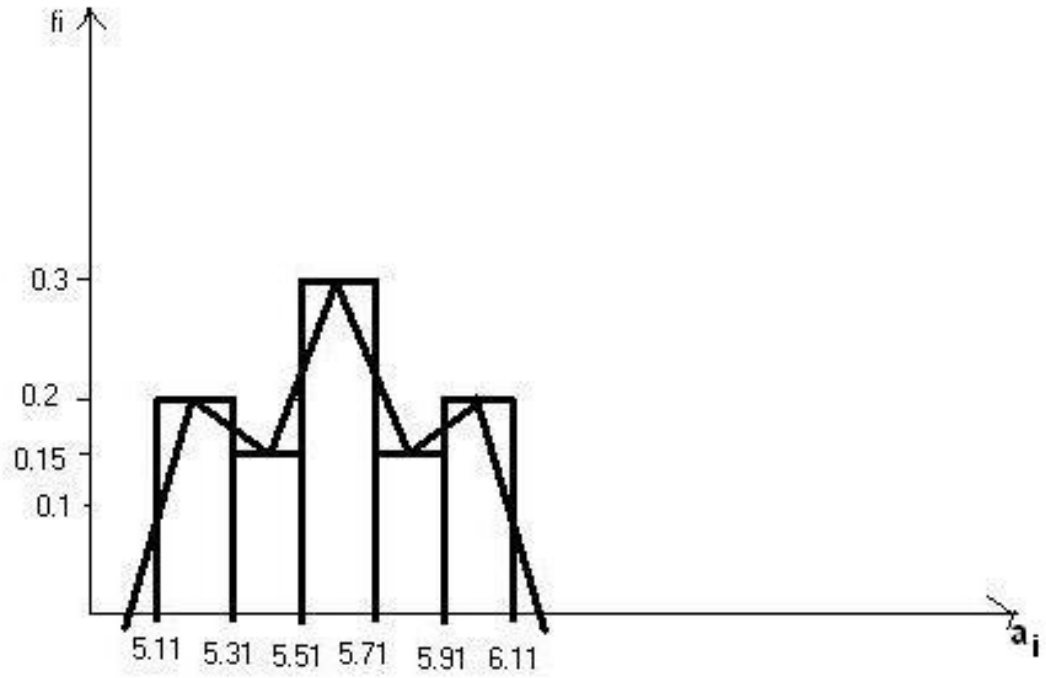


FIGURE 1.7 – Polygône des fréquences

1.3.3.3 Courbe des fréquences cumulées

La courbe des fréquences cumulées de la distribution $\{(]a_i,\, a_{i+1}],\, f_i)\,/\,1 \leq i \leq p\}$ s'obtient en joignant les points de coordonnées (a_{i+1}, F_i) où $F_i = f_1 + \cdots + f_i$, $i = 1, \cdots, p$ et $(x, 1)$ pour $x \geq a_{p+1}$.
Lorsque la borne inférieure de la première classe est observée c'est à dire l'intervalle est fermé en a_1, $F(a_1) \neq 0$, (comme c'est le cas dans l'exemple 1.3.6), on complète la courbe en joignant les points $(c_0,\, 0)$ et (a_2, F_1) où $c_0 = a_1 - \dfrac{a}{2}$ et $F_1 = f_1$.
Lorsque la borne inférieure de la première classe n'est pas observée c'est à dire l'intervalle est ouvert en a_1, $F(a_1) = 0$, on complète la courbe en joignant les points $(a_1,\, 0)$ et (a_2, F_1).

Exemple 1.3.9 *:* Représentation de la courbe des fréquences cumulées de la distribution de l'exemple 1.3.6.

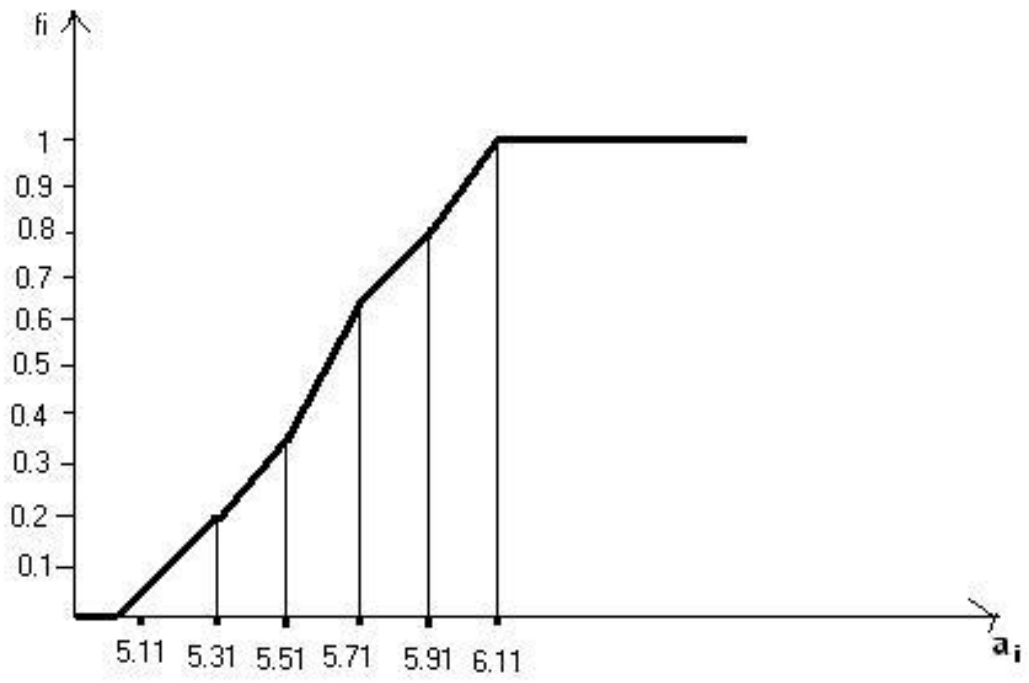


FIGURE 1.8 – Courbe des fréquences cumulées

1.4 Exercices

Exercice 1 : On a interrogé 30 sportifs sur leur degré de satisfaction face à leur entraineur. Les réponses possibles étaient : Très satisfait, satisfait, insatisfait et très insatisfait. Voici les données obtenue :

Très satisfait	satisfait	satisfait	satisfait	Insatisfait
Très satisfait	Insatisfait	satisfait	Très Insatisfait	satisfait
satisfait	Très satisfait	satisfait	Très satisfait	Insatisfait
Insatisfait	satisfait	Insatisfait	satisfait	satisfait
satisfait	Très Insatisfait	Très satisfait	satisfait	satisfait
satisfait	satisfait	satisfait	Très satisfait	Insatisfait

a) Quelle est la variable étudiée ? donner son type.

b) Présenter les données sous forme d'un tableau de distribution.

c) Présenter ces données sous forme de graphique.

Exercice 2 : Une garderie a compté pour chacun de ses enfants, combien de fois ils ont été absents de la garderie le mois de septembre 2012. Voici les résultats : 0 1 0 5 2 3 0 4 5 1 3 2 0 3 0 4 5 4 3 2 0 0 1 0 1 2 1 2 0 0 3 4 5 0 0 1 0 1 0 2 0 1 3 4 5 4 0 2 0 0.

a) Quelle est la variable étudiée ? donner son type.

b) Présenter les données sous forme d'un tableau de distribution.

c) Présenter ces données sous forme de graphique.

Exercice 3 : Examiner la série statistique suivante donnât la superficie occupée par un échantillon de 16 exploitation agricole au Maroc.

Ville	Superficie en hectare	Ville	Superficie en hectare
Settat	8.5	*Berkane*	11.7
ElJadida	15.4	*Rabat*	2.8
Marrakech	4.5	*Safi*	7.5
Agadir	13	*Berchid*	5.8
Beni − Mellal	11.2	*Taroudant*	4.8
Fès	9.6	*Mohammedia*	3.7
Casablanca	10.2	*Kénitra*	13.5
Meknés	8.1	*SidiKacem*	5.9

a) Quelle est la variable étudiée ? donner son type.

b) Présenter les données sous forme d'un tableau de distribution.

c) Présenter ces données sous forme de graphique.

Chapitre 2

Les mesures de tendance centrale et de dispersion

2.1 Les mesures de tendance centrale

La tendance centrale se propose de synthétiser l'ensemble d'une série statistique en faisant ressortir une position centrale de la valeur du caractère étudié. Il existe plusieurs mesures de tendance centrale.

Le mode , la médiane et la moyenne

2.1.1 Le mode

2.1.1.1 Variable qualitative ou quantitative discrète

Le mode est une valeur de la variable pour laquelle l'effectif ou la fréquence est maximal(e). Le mode est noté m_d. Une distribution peut être unimodale, bimodale ou plurimodale.

Exemple 2.1.1 *:* **i)** Considérons la distribution des notes d'un groupe d'étudiants.

x_i	8/20	9/20	10/20	11/20	12/20	13/02	14/20
n_i	2	7	12	17	11	6	3

l'effectif maximal est 17

La variable est quantitative discrète. On a $m_d = 11/20$. Cette distribution est unimodale.
ii) Considérons la distribution des couleurs des voitures dans un parking

x_i	Rouge	Blanche	Verte	Jaune	Noire	Grise
n_i	2	7	5	7	5	7

l'effectif maximal est 7

La variable est qualitative. Ici on a trois modes : Blanche, Jaune et Grise. Cette distribution est plurimodale.

2.1.1.2 Variable quantitative continue

Dans le cas d'une variable quantitative continue, les données sont regroupées en classes. Si les classes sont toutes de même amplitude, une classe modale est celle dont l'effectif ou la fréquence est le plus élevé(e).

Exemple 2.1.2 *:* Soit la distribution suivante

$[x_i,\ x_{i+1}[$	[500, 700[	[700, 900[	[900, 1100[	[1100, 1300]
f_i	0.21	0.34	0.25	0.2

la fréquence maximale est 0.34, donc la classe modale est [700, 900[.
le mode m_d (qui appartient à la classe modale) est :

$$m_d = x_{i+1} - a \times \frac{(f_i - f_{i+1})}{(f_i - f_{i+1}) + (f_i - f_{i-1})}$$

x_{i+1} est la borne supérieure de la classe modale, a l'amplitude, f_i la fréquence de la classe modale, f_{i-1} la fréquence de la classe qui précède la classe modale et f_{i+1} la fréquence de la classe qui suit la classe modale.

Application : La classe modale est [700, 900[car la fréquence la plus élevée est 0.34, $x_{i+1} = 900$, $a = 200$, $f_i = 0.34$, $f_{i-1} = 0.21$, $f_{i+1} = 0.25$ et

$$m_d = 900 - \frac{(0.34 - 0.25)}{(0.34 - 0.25) + (0.34 - 0.21)} = 818.1818182$$

Remarque : Si les classes ne sont pas de même amplitude, on doit obligatoirement corriger les effectifs et les fréquences (c'est à dire rendre les classes de même amplitude on prendra la plus petite amplitude) avant de :

$$\begin{cases} \textbf{Construire l'histogramme} \\ \textbf{Construire le polygône des fréquences} \\ \textbf{déterminer la classes modale} \end{cases}$$

Exemple 2.1.3 *:* Les salaires mensuels (en milliers de dirhams) du personnel d'une entreprise se répartissent comme suit :

Classe	Effectif n_i	fréquence f_i	fréquence cumulée $F(x_{i+1})$
$]2, 3]$	15	$0,19$	$0,19$
$]3, 4]$	20	$0,25$	$0,44$
$]4, 6]$	20	$0,25$	$0,69$
$]6, 10]$	24	$0,31$	1
Total	79	1	

Les classes ne sont pas de même amplitude, il faut donc corriger les données, la plus petite amplitude est $a = 1$

Classe	Effectif corrigé	fréquence
$[2, 3]$	15	$0, 19$
$]3, 4]$	20	$0, 25$
$]4, 5]$	10	$0, 125$
$]5, 6]$	10	$0, 125$
$]6, 7]$	6	$0, 0775$
$]7, 8]$	6	$0, 0775$
$]8, 9]$	6	$0, 0775$
$]9, 10]$	6	$0, 0775$
Total	79	1

Il est clair que $]3, 4]$ est la classe modale.
$x_{i+1} = 4$, $a = 1$, $f_i = 0.25$, $f_{i-1} = 019$, $f_{i+1} = 0.125$ et

$$m_d = 4 - \frac{(0.25 - 0.125)}{(0.25 - 0.125) + (0.25 - 0.19)} = 3.324324324$$

2.1.2 La médiane

La médiane est la valeur m qui partage les éléments de la série statistique, préalablement classés par ordre croissant, en deux groupes d'effectifs égaux : 50% des individus présentent une valeur inférieure ou égale à la médiane et 50% présentent une valeur supérieure ou égale à la médiane.

2.1.2.1 Variable quantitative discrète

Soient $x_1, x_2, \cdots, x_N$ les valeurs prises par la variable. On les ordonne de la plus petite à la plus grande et on note $x_{(1)}$ la plus petite valeur $x_{(2)}$ la deuxième valeur, $\cdots$, $x_{(i)}$ la i^{me} valeur, $\cdots$ $x_{(N)}$ la plus grande valeur. Alors on a

$$m = \begin{cases} x_{(\frac{N+1}{2})} & \text{si } N \text{ est impair} \\ \dfrac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} & \text{si } N \text{ est pair} \end{cases}$$

Exemple 2.1.4 *:* **i)** Considérons la distribution suivante

x_i	10	20	30	40	50	60
n_i	3	8	4	9	3	3
effectifs cumulés	3	11	15	24	27	30

On a $N = 30$

donc N est pair d'où $\frac{N}{2} = 15$ et $m = \frac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} = \frac{x_{(15)} + x_{(16)}}{2} = \frac{30 + 40}{2} = 35$.
$x_{(16)} = 40$ car le premier effectif cumulé supérieur ou égal à 16 est 24 et $x_{(24)} = 40$.

ii) Considérons la distribution suivante

x_i	10	20	30	40	50	60
n_i	4	9	4	8	3	3
effectifs cumulés	4	13	17	25	28	31

On a $N = 31$

donc N est impair d'où $\frac{N+1}{2} = 16$ et $m = x_{(16)} = 30$ car le premier effectif cumulé supérieur ou égal à 16 est 17 et $x_{(17)} = 30$.

2.1.2.2 Variable quanitative continue

La médiane est la solution de l'équation $F(x) = 0,5$. Pour la déterminer, on commence par déterminer la classe médiane $]x_i, x_{i+1}]$ qui vérifie

$$F(x_i) \leq 0,5 \quad \text{et} \quad F(x_{i+1}) \geq 0,5$$

La médine m (qui appartient à la classe médiane) est

$$m = x_i + (x_{i+1} - x_i)\frac{0.5 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

x_{i+1} est la borne supérieure de la classe médiane, x_i la borne inférieure de la classe médiane, $F(x_{i+1})$ la fréquence de la classe médiane et $F(x_i)$ la fréquence de la classe qui précède la classe médiane.

Exemple 2.1.5 *:* Reprenons l'exemple de la distribution des salaires mensuels (en milliers de dirhams) du personnel d'une entreprise :

Classe	Effectif n_i	fréquence f_i	fréquence cumulée $F(x_{i+1})$
$]2, 3]$	15	$0,19$	$0,19$
$]3, 4]$	20	$0,25$	$0,44$
$]4, 6]$	20	$0,25$	$0,69$
$]6, 10]$	24	$0,31$	1
Total	79	1	

il est clair que la classe médiane est]4, 6].

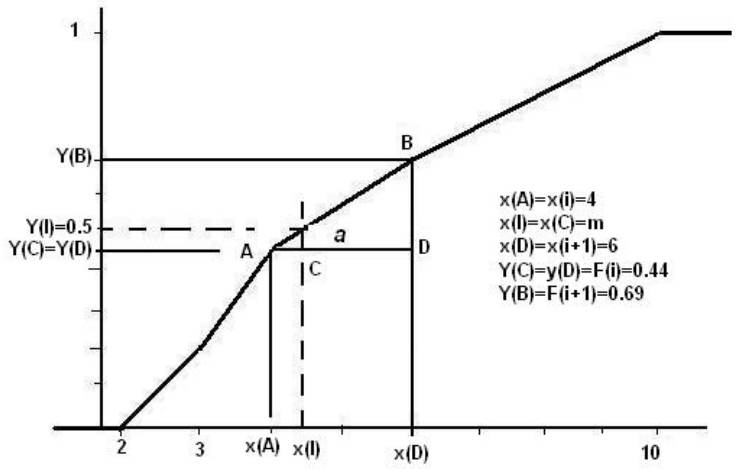


FIGURE 2.1 – Courbe des fréquences cumulées

On a $F(4) = 0,44 < 0.5$ et $F(6) = 0.64 > 0.5$, la classe médiane est donc $]4, 6]$.
$x_i = 4$, $x_{i+1} = 6$, $F(x_i) = 0.44$, $F(x_{i+1}) = 0.69$ et

$$m = 4 + (6 - 4)\frac{0,5 - 0,44}{0,69 - 0,44} = 4,48$$

2.1.3 La Moyenne arithmétique

2.1.3.1 Variable quantitative discrète

La moyenne arithmétique notée $\overline{x}$, est égale à la somme des valeurs distinctes de la variable multipliées par leurs effectifs respectifs divisée par la somme des effectifs.

$$\overline{x} = \frac{\sum\limits_i n_i x_i}{\sum\limits_i n_i} = \frac{\sum\limits_i n_i x_i}{N}$$

et comme $f_i = \dfrac{n_i}{N}$ on a aussi $\overline{x} = \sum\limits_i f_i x_i$

Exemple 2.1.6 *:* Considérons la distribution de l'exemple 2.1.4 i)

$$\overline{x} = \frac{10 \times 4 + 20 \times 8 + 30 \times 4 + 40 \times 9 + 50 \times 3 + 60 \times 3}{4 + 8 + 4 + 9 + 3 + 3} = \frac{1010}{31} = 32.58064516$$

2.1.3.2 Variable quantitative continue

La moyenne arithmétique notée toujours $\overline{x}$, est égale à la somme des centres des classes de la variable multipliées par leurs effectifs respectifs divisée par la somme des effectifs.

$$\overline{x} = \frac{\sum\limits_i n_i c_i}{\sum\limits_i n_i} = \frac{\sum\limits_i n_i c_i}{N}$$

où c_i est le centre de de la classe associée à l'effectif n_i.
et comme $f_i = \dfrac{n_i}{N}$ on a aussi $\overline{x} = \sum_i f_i c_i$

Exemple 2.1.7 *:* Reprenons l'exemple de la distribution des salaires mensuels

Classe	Effectif n_i	fréquence f_i	fréquence cumulée $F(x_{i+1})$
$]2,3]$	15	$0,19$	$0,19$
$]3,4]$	20	$0,25$	$0,44$
$]4,6]$	20	$0,25$	$0,69$
$]6,10]$	24	$0,31$	1
Total	79	1	

$$\overline{x} = \frac{15 \times 2,5 + 20 \times 3,5 + 20 \times 5 + 24 \times 8}{15 + 20 + 20 + 24} = \frac{399,5}{79} = 5,05$$

2.2 Les mesures de dispersion

Les indicateurs de dispersion sont nombreux, les plus courants sont :
L'étendue, l'intervalle interquartile, la variance et le cœfficient de variation.

2.2.1 L'étendue

2.2.1.1 Variable quantitative discrète

L'étendue mesure l'écart entre la plus petite valeur de la variable et la plus grande :

$$e = x_{max} - x_{min}$$

où x_{min} (resp. x_{max}) est la valeur minimale (resp. maximale) prises par la variable.

Exemple 2.2.1 *:* Soient les 4 series statistiques suivantes
a) 10, 10, 10, 10, 20, 30, 30, 30, 30 $\quad \overline{x} = \dfrac{4 \times 10 + 1 \times 20 + 4 \times 30}{9} = \dfrac{180}{9} = 20$
b) 20, 22, 21, 20, 20, 19, 18, 20, 20 $\quad \overline{x} = \dfrac{18 + 19 + 5 \times 20 + 21 + 22}{9} = \dfrac{180}{9} = 20$
c) 1, 4, 6, 8, 20, 32, 34, 36, 39 $\quad \overline{x} = \dfrac{1 + 4 + 6 + 8 + 20 + 32 + 34 + 36 + 39}{9} = \dfrac{180}{9} = 20$
d) 10, 12, 14, 16, 20, 24, 26, 28, 30 $\quad \overline{x} = \dfrac{10 + 12 + 14 + 16 + 20 + 24 + 26 + 28 + 30}{9} = \dfrac{180}{9} = 20$
Ces quatre séries ont la même moyenne $\overline{x} = 20$ et la même médiane $m = 20$. Pourtant ces séries sont trés différentes. Cette différence provient de leur dispersion, en effet :
$Etendue(a) = 30 - 10 = 20$; $Etendue(b) = 22 - 18 = 4$; $Etendue(c) = 39 - 1 = 38$; $Etendue(d) = 30 - 10 = 20$.
Quoique les séries a) et d) ont la même étendue, les valeurs de la série d) contrairement à celles de la série a) sont uniformément espacées.

2.2.1.2 Variable quantitative continue

Dans ce cas l'étendue est la différence entre les centres des classes extrêmes

$$e = c_{max} - c_{min}$$

2.2.2 Les quartiles, les déciles et les centiles

Nous savons que la médiane divise la distribution en deux parties égales. Il existe d'autres indicateurs utiles :

a) Les quartiles qui divise la distribution en quatre (4) parties égales

b) Les déciles qui divise la distribution en dix (10) parties égales

c) Les ceniles qui divise la distribution en cent (100) parties égales

Les quartiles sont notés Q_1, Q_2 et Q_3 et on a $F(Q_1) = 0.25$, $F(Q_2) = 0.5$ et $F(Q_3) = 0.75$. La médiane est le 2^{eme} quartile, le 5^{eme} décile et le 50^{eme} centile. Des techniques similaires à celles utilisées pour déterminer la médiane permettent de déterminer ces indicateurs.
Pour le premier quartile

$$\left.\begin{array}{c} x_i \leq Q_1 \leq x_{i+1} \\ F(x_i) \leq 0,25 \leq F(x_{i+1}) \end{array}\right\} \text{ et } Q_1 = x_i + (x_{i+1} - x_i)\frac{0,25 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

Pour le troisième quartile

$$\left.\begin{array}{c} x_i \leq Q_3 \leq x_{i+1} \\ F(x_i) \leq 0,75 \leq F(x_{i+1}) \end{array}\right\} \text{ et } Q_3 = x_i + (x_{i+1} - x_i)\frac{0,75 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

Exemple 2.2.2 *:* Reprenons la distribution des salaires mensuels.

Classe	Effectif n_i	fréquence f_i	fréquence cumulée $F(x_{i+1})$
$]2,3]$	15	$0,19$	$0,19$
$]3,4]$	20	$0,25$	$0,44$
$]4,6]$	20	$0,25$	$0,69$
$]6,10]$	24	$0,31$	1
Total	79	1	

$0.19 \leq F(Q_1) = 0.25 \leq 0.44 \Longrightarrow 3 \leq Q_1 \leq 4$, d'où $Q_1 = 3 + (4-3) \times \dfrac{0,25 - 0,19}{0,44 - 0,19} = 3,24$

$0.69 \leq F(Q_3) = 0.75 \leq 1 \Longrightarrow 6 \leq Q_3 \leq 10$, d'où $Q_3 = 6 + (10-6) \times \dfrac{0,75 - 0,69}{1 - 0,69} = 6,19.$

2.2.3 L"intervalle interquartile

Q_1 étant le premier quartile et Q_3 étant le troisième quartile, l'intervalle $[Q_1, Q_3]$ est appelé intervalle interquartile . Il contient 50% des observations, le reste se réparti avec

25% à gauche de Q_1 et 25% à droite de Q_3. On note $\mathbf{R}(Q_3 - Q_1)$ la largeur de l'intervalle interquartile.

Exemple 2.2.3 *:* Reprenons l'exemple de la distribution des salaires mensuels.
L'intervalle $[3,24; 6,19]$ est l'intervalle interquartile sa largeur est

$$\mathbf{R}(Q_3 - Q_1) = 6,19 - 3,24 = 2,85$$

2.2.4 La variance et l'écart-type

La variance est un résumé statistique qui mesure la concentration ou la dispersion des observations autour de la moyenne

2.2.4.1 Variable quantitative discrète

La variance $V(x)$ est la moyenne arithmétique des carrés des écarts des valeurs de la variable à la moyenne arithmétique

$$V(x) = \frac{1}{N}\sum_i n_i(x_i - \overline{x})^2 = \sum_i f_i(x_i - \overline{x})^2 \quad \text{où } N = \sum_i n_i$$

La racine carrée de la variance est appelée l'écart-type

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_i n_i(x_i - \overline{x})^2} = \sqrt{\sum_i f_i(x_i - \overline{x})^2}$$

Exemple 2.2.4 *:* Considérons la distribution suivante

x_i	10	20	30	40	50	60
n_i	4	8	4	9	3	3

on a $N = 31$ et $\overline{x} = 32.58$

$$\begin{aligned} V(x) &= \frac{4(10-32.58)^2 + 8(20-32.58)^2 + 4(30-32.58)^2}{31} \\ &\quad + \frac{9(40-32.58)^2 + 3(50-32.58)^2 + 3(60-32.58)^2}{31} \\ &= \frac{6993.5484}{31} = 225.598 \\ \sigma(x) &= \sqrt{225.598} = 15.02 \end{aligned}$$

Relation de König : $\sum_i n_i(x_i - \overline{x})^2 = \sum_i n_i x_i^2 - N\overline{x}^2 \Longrightarrow V(x) = \frac{1}{N}\left(\sum_i n_i x_i^2\right) - \overline{x}^2$

2.2.4.2 Variable quantitative continue

La variance $V(x)$ est la moyenne arithmétique des carrés des écarts des centres des classes à la moyenne arithmétique

$$V(x) = \frac{1}{N}\sum_i n_i(c_i - \overline{x})^2 = \sum_i f_i(c_i - \overline{x})^2 \quad \text{où } c_i \text{ est le centre de la classe associée à } n_i$$

La racine carrée de la variance est appelée l'écart-type

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_i n_i(c_i - \overline{x})^2} = \sqrt{\sum_i f_i(c_i - \overline{x})^2}$$

Exemple 2.2.5 *:*Reprenons la distribution des salaires mensuels.

Classe	Effectif n_i	fréquence f_i	fréquence cumulée $F(x_{i+1})$
$]2,3]$	15	$0,19$	$0,19$
$]3,4]$	20	$0,25$	$0,44$
$]4,6]$	20	$0,25$	$0,69$
$]6,10]$	24	$0,31$	1
Total	79	1	

on a $\overline{x} = 5.05$

$$\begin{aligned} V(x) &= \frac{15(2.5-5.05)^2 + 20(3.5-5.05)^2 + 20(5-5.05)^2}{79} \\ &= +\frac{24(8-5.05)^2}{79} = \frac{354.497}{79} = 4.487 \\ \sigma(x) &= \sqrt{4.487} = 2.2.118 \end{aligned}$$

Relation de König : $\sum_i n_i(c_i - \overline{x})^2 = \sum_i n_i c_i^2 - N\overline{x}^2 \Longrightarrow V(x) = \frac{1}{N}\left(\sum_i n_i c_i^2\right) - \overline{x}^2$

Propriété 2.2.1 *:* Soient $x_i, \cdots, x_N$ une série statistiques et a, b deux réels, alors $V(x+b) = V(x)$, $V(a\,x) = a^2V(x)$, $V(a\,x+b) = V(a\,x) = a^2V(x)$ et $\sigma(ax) = |a|\sigma(x)$

2.2.5 Le cœfficient de variation

Tous les indicateurs de dispersion que nous avons vu jusqu'à présent dépendent des unités de mesure de la variable. Ils ne permettent pas de comparer des dispersions de distributions statistiques hétérogènes.
Le cœfficient de variation, qui est un nombre sans dimension, permet cette comparaison lorsque les valeurs de la variable sont positives. Il s'écrit

$$CV = \frac{\sigma(x)}{\overline{x}}$$

Si $CV < 0,5$ alors la dispersion n'est pas importante. Si $CV > 0,5$ alors la dispersion est importante.

Exemple 2.2.6 *:* Dans une maternité on a relevé le poids (en kg) à la naissance de 47 nouveau-nés. Les données collectées sont résumées dans le tableau suivant :

classe	n_i	c_i	$n_i c_i$	$(c_i - \overline{x})$	$(c_i - \overline{x})^2$	$n_i(c_i - \overline{x})^2$
$]2,5;3,0]$	8	$2,75$	$22,00$	$-0,73$	$0,5329$	$4,2632$
$]3,0;3,5]$	15	$3,25$	$48,75$	$-0,23$	$0,0529$	$0,7935$
$]3,5;4,0]$	20	$3,75$	$75,00$	$0,27$	$0,0729$	$1,4580$
$]4,0;4,5]$	4	$4,50$	$18,00$	$0,52$	$0,2704$	$1,0816$
Total	47		$163,75$			$7,5963$

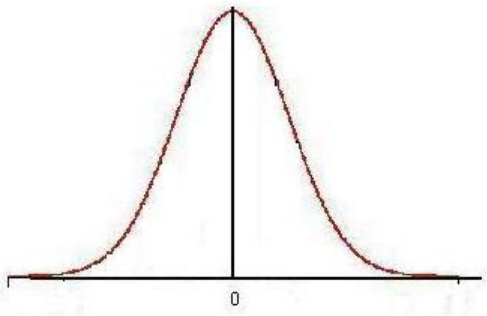
$\overline{x} = \dfrac{163,75}{47} = 3,48$, $\sigma(x) = \sqrt{\dfrac{7,5963}{47}} = \sqrt{0,1616} = 0,4019$ et $CV = \dfrac{0,4019}{3,48} = 0,1154$
Le cœfficient de variation étant faible, le poids à la naissance est concentré autour de la moyenne.

2.3 Symétrie et asymétrie

2.3.1 Symétrie et asymétrie

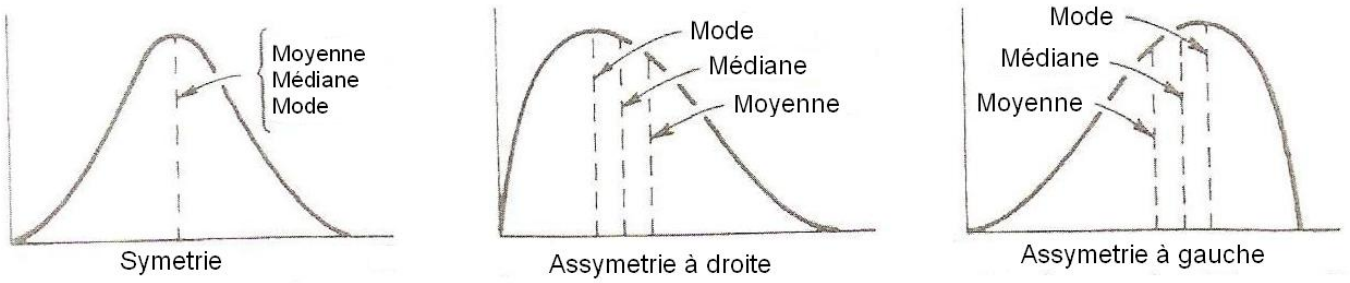
Une distribution est dite symétrique si le mode, la médiane et la moyenne sont confondus. Une distribution qui n'est pas symétrique est dite asymétrique

Remarque : Une variable statistique est symétrique si ses valeurs sont réparties de manière symétrique autour de la moyenne c'est à dire si le polygône des fréquences a la forme d'une clôche comme dans la figure ci-après.



A la différence de la médiane, la moyenne arithmétique est fortement influencée par les valeurs extrêmes. Lorsque les valeurs sont distribuées de manière symétrique, la moyenne arithmétique coïncide avec la médiane. Lorsque la distribution est asymétrique, la moyenne

arithmétique dépasse la médiane si les valeurs extrêmes sont élevées et se situe en dessous de la médiane si les valeurs extrêmes sont basses.
Une distribution est dite asymétrique à droite, si la courbe du polygône des fréquences est étalée à droite, on a généralement : mode $<$ médiane $<$ moyenne.
Une distribution est dite asymétrique à gauche, si la courbe du polygône des fréquences est étalée à gauche, on a généralement : moyenne $<$ médiane $<$ mode.
La figure ci-dessous illustre ces diffférents cas.



2.3.2 Coefficient de dissymétrie

le coefficient de dissymétrie D_P dit de Pearson a pour rôle de fournir une mesure de dissymétrie d'une distribution. La mesure de dissymétrie est calculée à partir de la formule suivante :

$$D_P = \frac{3 \times (\overline{X} - m)}{\sigma}$$ où $\overline{X}$ est la moyenne, m la médiane et σ l'écart type.

- Si la distrbution est symétrique, D_P est nulle car $\overline{X} = m$.
- Si la courbe est assymétrique à droite, la courbe est étalée à droite et D_P est positif car $\overline{X} > m$.
- Si la distribution est assymétrique à gauche, la courbe est étalée à gauche et D_P est négatif car $\overline{X} < m$.

2.4 Applications : Le théorème de Tchebychebv

Nous avons vu qu'il existe plusieurs mesures de positions et de dispersions. La moyenne est sans doute la mesure de position la plus répandue alors que la variance et l'écart-type sont les mesures de dispersion les plus utilisées. Nous allons voir comment en n'utilisant que la moyenne et l'écart-type, il est possible d'explorer des données.
Le théorème de Tchebychebv permet d'évaluer le pourcentage des données qui se trouvent

à k écart-types de la moyenne cest à dire le pourcentage des données appartenant à l'intervalle $[\overline{x} - k\,\sigma\,,\, \overline{x} + k\,\sigma]$, pour un entier k donné.

Théorème 2.4.1 *:* Pour un entier $k \geq 2$, au moins $100 \times (1 - \frac{1}{k^2})\%$ des observations, d'une série de données, se trouvent à k écart-type de la moyenne de cette série.

Exemple 2.4.1 *:*
Les notes de 100 étudiants d'un contrôle de statistique ont une moyenne $\overline{x} = 14$ avec un écart-type $\sigma(x) = 1$. combien d'étudiants ont une note entre 12 et 16 ?
Remarquons que $12 = \overline{x} - 2\sigma(x)$ et que $16 = \overline{x} + 2\sigma(x)$. Ainsi, d'après le théorème de Tchebychev, le pourcentage d'étudiants ayant obtenue une note entre 12 et 16 est supérieur ou égal à $100 \times (1 - \frac{1}{2^2})\% = 75\%$.
Le pourcentage garanti par le théorème de Tchebychev peut être améliorer sous certaines conditions.

Règle Empirique
Si les observations sont réparties de manière symétrique autour de la moyenne alors
• Approximativement 68% des valeurs sont à un écart-type de la moyenne.
• Approximativement 95% des valeurs sont à deux écart-type de la moyenne.
• Approximativement toutes les valeurs sont à trois écart-type de la moyenne.

2.5 Exercices

Exercice 1 : La répartition des célibataires selon leur âge est fournie par le tableau :

Age	]15, 30]	]30, 40]	]40, 50]	]50, 60]	]60, 70]	]70, 80]	]80, 90]
n_i	4500	450	400	230	200	?	20

1) Sachant que l'âge moyen est égal à 28,8038 ans, déterminer l'effectif manquant.
2) Qu'elle est la proportion des célibataires dont l'âge est
a) inférieur ou égal à 40 ?, **b)** supérieur à 50 ?, **c)** Compris entre 30 et 60 ?
3) Déterminer l'âge médian.
4) Sachant que 75% des célibataires ont un âge inférieur à x, déterminer x.

Exercice 2 : La moyenne semestrielle des notes (de 0 à 20) d'une classe d'éléves de terminale est de 8,5 et leur écart-type est de 2,5. Il n'ya pas de notes supérieures à 18. Le professeur veut changer les notes afin d'obtenir un moyenne égale à 10 et un écart-type égal à 2.
On note x l'ancienne note et y la nouvelle. Le professeur utilise la transformation $y = ax+b$ où $a > 0$ et b sont deux nombres réels. Déterminer a et b et vérifier que ce changement est possible.
Aprés le changement des notes, donner le pourcentages d'éléves qui ont une note comprise entre 6 et 14.

Chapitre 3

Liaisons entre deux variables statistiques

L'étude statistique peut se porter sur deux caractères présents dans tous les membres de la population. Ces deux caractères sont représentés par deux variables X et Y. On peut utiliser l'information dont on dispose pour étudier la liaison qui existe éventuellement entre ces deux caractères.

3.1 Représentation graphique du nuage de points

Une étude simultanée sur deux variables quantitatives X et Y sur une population de n individus a donné les différents points de mesures :

$$(x_1,\, y_1),\, (x_2,\, y_2),\, (x_3,\, y_3), \cdots,\, (x_{n-1},\, y_{n-1}), (x_n,\, y_n)$$

Ces données sont représentées par paires. le premier élément de la paire correspond à la valeur prise par la variable X et le second par Y.
On représente une distribution statistique à deux caractères quantitatifs par l'ensemble des points A_k, $k = 1 \cdots$, n de coordonnées $\{(x_k,\, y_k),\, k = 1 \cdots n\}$, chaque individu correspond à un point du plan.
On appelle nuage de points l'ensemble des points A_k, $k = 1 \cdots$, n de coordonnées $\{(x_k,\, y_k),\, k = 1 \cdots n\}$. La représentation graphique du nuage de points est essentielle pour déterminer s'il existe ou non une relation entre les variables X et Y.
On représente sur l'axe des abscisse les mesures x_k, $k = 1 \cdots$, n et sur l'axe des ordonnées les mesures y_k, $k = 1 \cdots$, n est le points A_k correspond à la paire $(x_k,\, y_k)$.

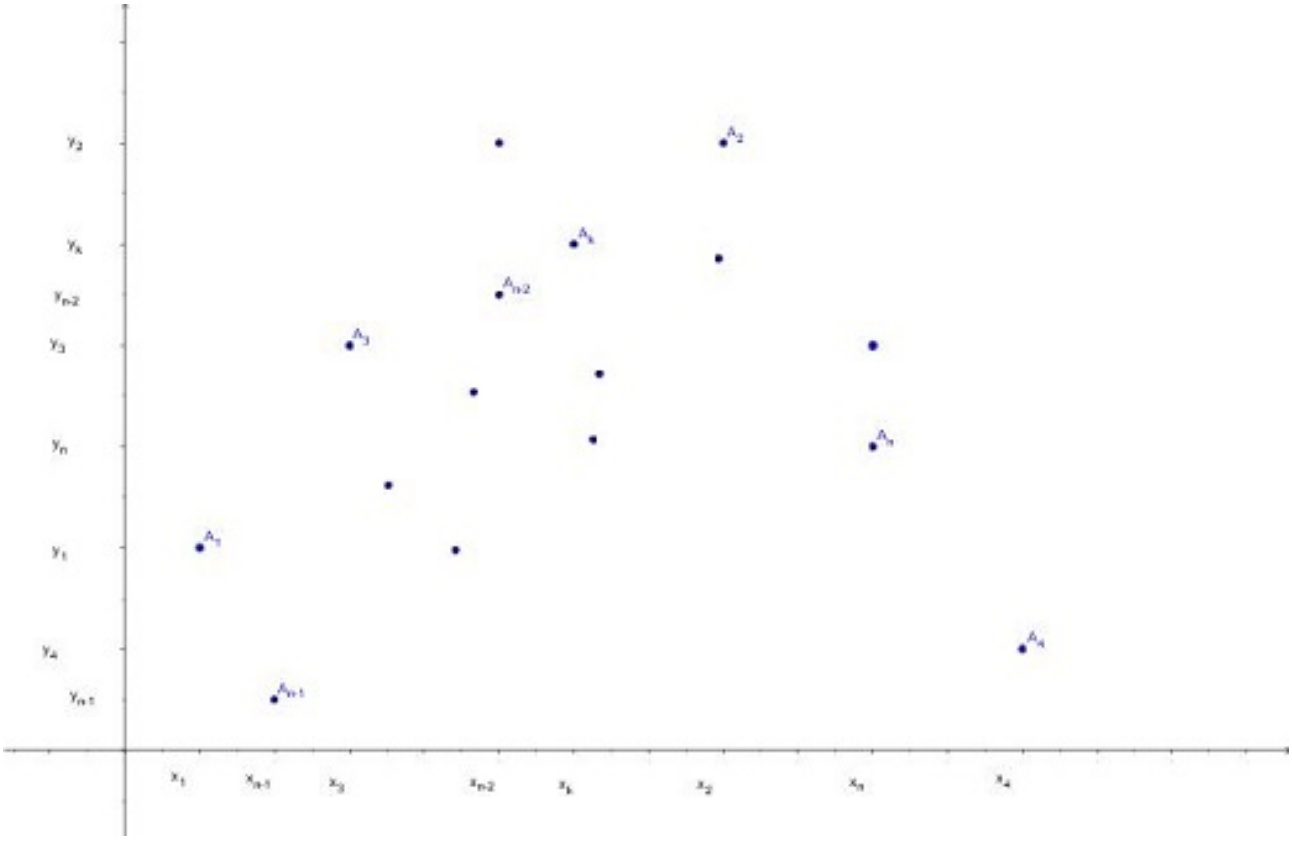


FIGURE 3.1 – Nuage de points

3.2 Ajustement linéaire

L'objectif est de mettre en évidence l'existence d'une relation entre deux variables quantitatives (continues ou discrètes.)

3.2.1 Covariance et coefficient de corrélation

La covariance des variables X et Y s'écrit :

$$Cov(x,\, y) = \frac{1}{N}\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{x})$$

La covariance dépend des unités de mesures dans lesquelles sont exprimées les variables. De même, on définit le coefficient de corrélation :

$$\rho_{xy} = \frac{Cov(x,\, y)}{\sigma(x)\,\sigma(y)}$$

qui est un nombre sans dimension destiné à mesurer l'intensité de la liaison entre les variations de la variable X et celles de Y.

On a toujours :

$$-1 \leq \rho_{xy} \leq 1$$

Si $|\rho_{xy}| = 1$ les points (x_k, y_k), $k = 1 \cdots, n$ sont alignés, alors il existe une liaison linéaire entre X et Y c'est à dire, il existe deux réels a et b tel que

$$Y = a\,X + b$$

Si $\rho_{xy} = 0$ les variables X et Y sont non corrélées linéairement c'est à dire il n'existe pas de liaison linéaire entre X et Y.
En pratique si $|\rho_{xy}|$ est proche de 1, on dit qu'il y a corrélation linéaire entre les variables X et Y. La corrélation est d'autant plus forte que $|\rho_{xy}|$ est proche de 1.

Exemple 3.2.1 *:* Considérons dans une entreprise, la variable X : les dépenses en milliers de dirhams en publicité et Y : les ventes en milliers de dirhams des articles produit.

$x_i \times 1000DH$	$y_i \times 1000DH$	$x_i \times y_i$	x_i^2	y_i^2
1.7	50	85	2.89	2500
3.0	100	300	9	1000
2.0	75	150	4	5625
1.5	45	67.50	2.25	2025
0.6	20	12	0.36	400
1.5	50	75	2.25	2500
10.3	340	689.50	20.75	23050

$$\overline{x} = \tfrac{10.3}{6} = 1.717 \qquad \overline{y} = \tfrac{340}{6} = 56.667$$
$$V(x) = \tfrac{20.75}{6} - 1.717^2 = 0.51 \qquad V(y) = \tfrac{23050}{6} - 56.667^2 = 630.52$$
$$Cov(x,\, y) = \tfrac{689.50}{6} - 1.717 \times 56.667 = 17.62 \qquad \rho_{xy} = \tfrac{17.62}{0.714 \times 25.11} = 0.98$$

Le coefficient de corrélation étant proche de 1 on peut conclure que les ventes augmentent en même temps que les dépenses de publicité.

3.2.2 Droite de régression

Si ρ_{xy} est proche de 1 ($|\rho_{xy} > 0.8|$ et si l'examen du nuage de points indique qu'on peut supposer une relation de type linéaire entre X et Y, alors on cherche à déterminer les réels a et b de la droite

$$Y = a\,X + b$$

telle que la distance entre cette droite et chaque point du nuage soit le plus petit possible. La méthode des moindres carrés propose cette notion de proximité entre la droite et le nuage des points. elle consiste à minimiser la fonction

$$\phi(a,\, b) = \sum_{k=1}^{n} (y_k - a\,x_k - b)^2$$

si on note $\overline{x}$ et $\overline{y}$ les moyennes respectives de X et Y, alors le couple $(\hat{a},\, \hat{b})$ qui minimise la fonction ϕ est

$$\begin{cases} \hat{a} &= \dfrac{Cov(x,\, y)}{V(x)} \\ \hat{b} &= \overline{y} - a\,\overline{x} \end{cases}$$

La droite $y = a\,x + b$ est appelée droite de régression linéaire.

Remarque : La droite de régression $y = a\,x + b$ passe par les points $(0,\, b)$ et $(\overline{x},\, \overline{y})$.

Exemple 3.2.2 :On dispose des mesures de taille en cm (variable X) et de poids en kg (variable Y) de 20 enfants d'une école.

	1	2	3	4	5	6	7	8	9	10
X	132	132	131	128	133	125	133	128	129	126
Y	24.75	24.55	22.5	21.46	25.92	24.15	27.86	28.34	25.82	28.5
	11	12	13	14	15	16	17	18	19	20
X	139	135	140	136	134	137	142	143	141	135
Y	33.11	33.89	33.88	29.07	31.61	30.68	40.51	35.45	35.11	31.27

$$\overline{x} = \frac{2679}{20} = 133.95 \quad \overline{y} = \frac{588.43}{20} = 29.42$$

$$V(x) = \frac{530.95}{20} = 26.55 \quad V(y) = \frac{469.3}{20} = 23.47$$

$$Cov(x,\, y) = \frac{409.36}{20} = 20.47 \quad \rho_{xy} = \frac{20.47}{\sqrt{26.55 \times 23.47}} = 0.82$$

$\rho_{xy} = 0.82 > 0.8$ donc il existe a et b telque $Y = a\,X + b$ avec

$$a = \frac{Cov(x,\, y)}{V(x)} = \frac{20.47}{26.55} = 0.77 \;\;,\;\; b = \overline{y} - a\,\overline{x} = 29.42 - 0.77 \times 133.95 = -73.72$$

La doite de régression est $y = 0.77 \times x - 73.72$ elle passe par les points $(0,\, -73.72)$, $(133.95,\, 29.42)$.

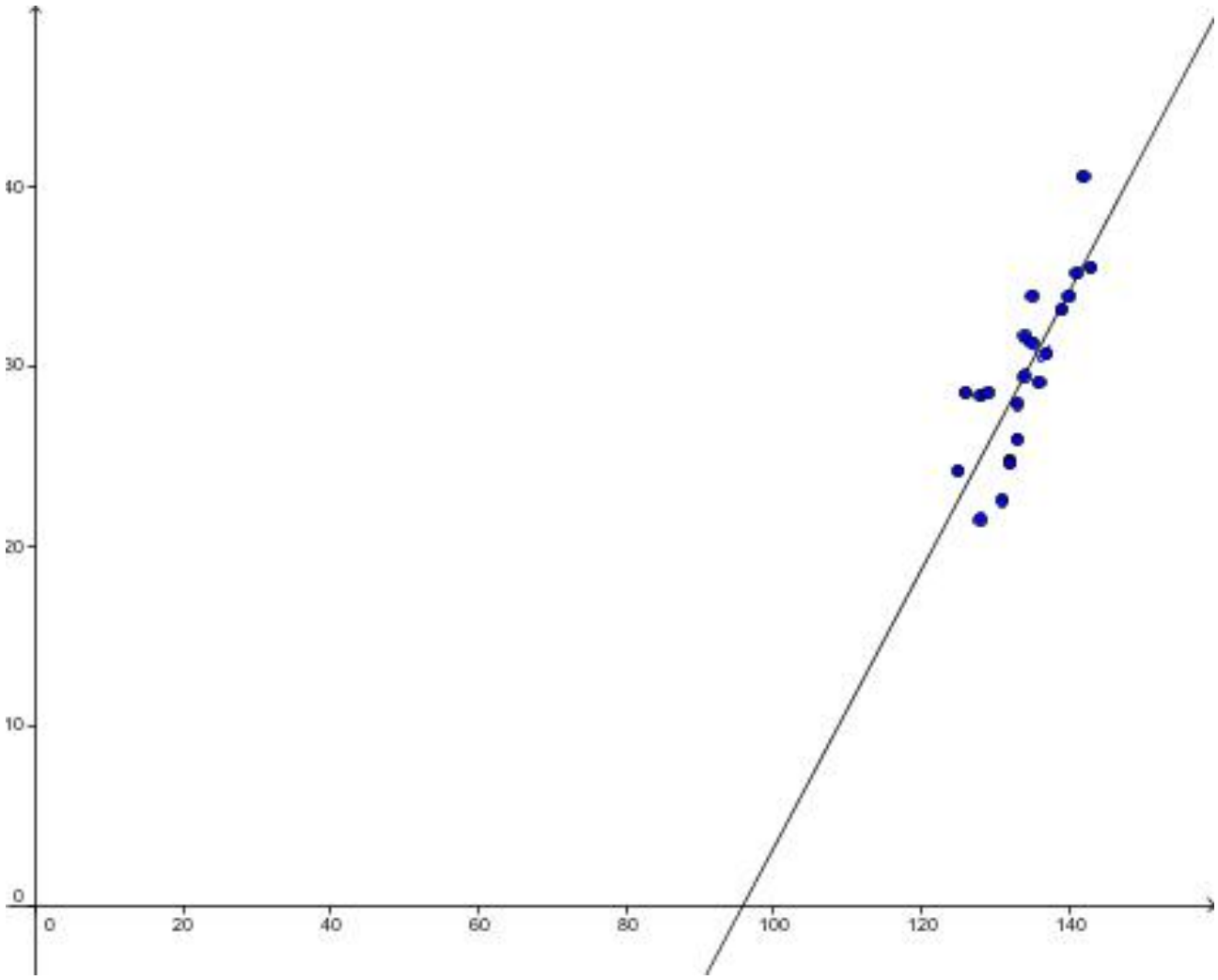


FIGURE 3.2 – nuage de points et droite de régression

3.3 Exercice

Exercice : Une entreprise souhaite étudier comment varie le coût annuel (en centaine de DHs de maintenance d'un véhicule utilitaire d'un type donné en fonction de son âge (en mois). Nous disposons des données suivantes :

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Coût annuel	48	43	77	89	50	40	56	62	100	47	71	58	102	35	60
Âge	15	8	36	41	16	8	21	21	53	10	32	17	58	6	20

Nous notons y_i le coût annuel de maintenance pour le véhicule utilitaire i et x_i son âge.
1) Calcler la covariance entre X et Y ainsi que la coefficient de corrélattion
2) peut-on ajuster y par x ? justifier la réponse. Si oui, calculer a et b les paramètres de la droite de régression $y = a\,x + b$ et représenter le nuage de points et la droite de régression.