Université Mohammed V
Faculté des Sciences
Département de Mathématiques.
Rabat Maroc

Filière : SMI

Semestre 3

Module 18

Cours de Statistique Descriptive

Par le

Professeur HAKAM Samir

Année : 2016 - 2017

Table des matières

Introduction

La statistique désigne l'ensemble des méthodes mathématiques relative à la collecte, à la présentation, à l'analyse et à l'utilisation des données numériques. Ces opérations permettent de tirer des conclusions et de prendre des décisions dans les situations d'incertitudes qu'on rencontre dans les domaines scientifiques, économiques, sciences soiales ou des affaires ...

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 *:*

i) La couleur des cheveux, Etat du temps constaté à Rabat pendant le mois de Janvier 2015 (pluvieux, orageux, beau, venteux, brouillard, ...), mode de transport pour se rendre à la faculté (voiture, taxi, bus, tramway, moto, bicyclette, à pied) définissent des variables qualitatives.

ii) La taille, le poids, le salaire, l'âge, les températures matinales relevées sous abri chaque jour

à Rabat, les notes sur 20 obtenues en statistique par les étudiants SMI, la hauteur des précipitations tombées chaque mois à Rabat 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

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

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 voitures 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.

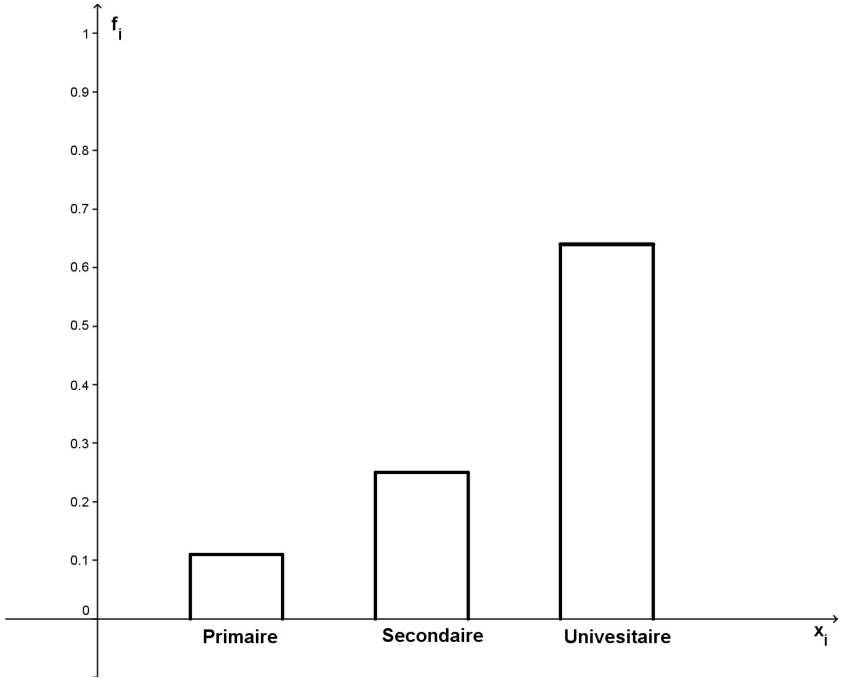

FIGURE 1.1 – Diagramme en tuyaux d'orgues

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.

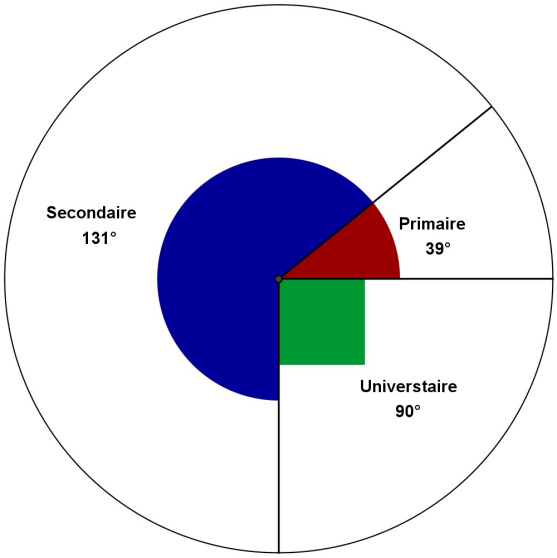

FIGURE 1.2 – Diagramme circulaire

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.

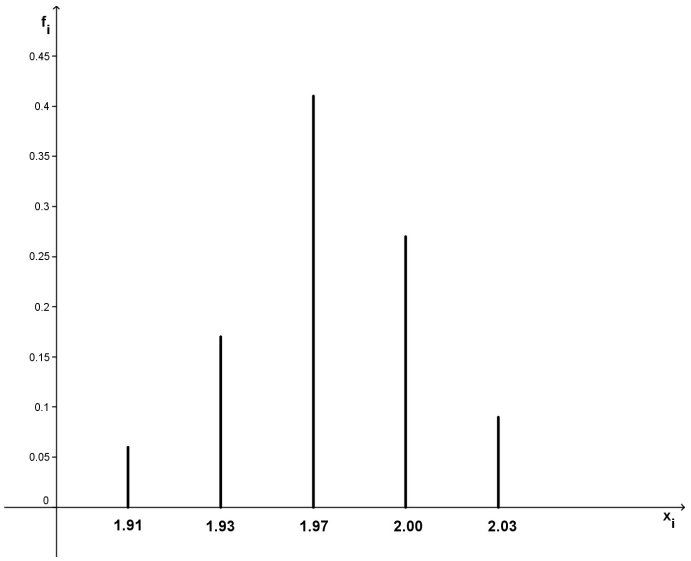

FIGURE 1.3 – Diagramme en bâtons

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.

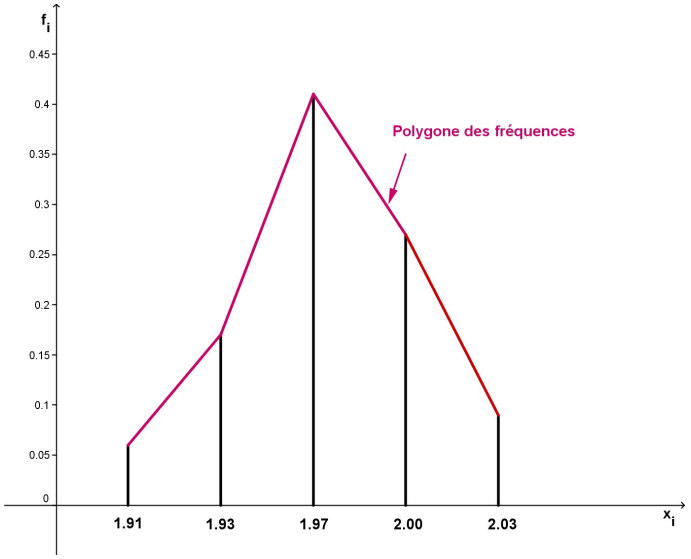

FIGURE 1.4 – Polygône des fréquences

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.

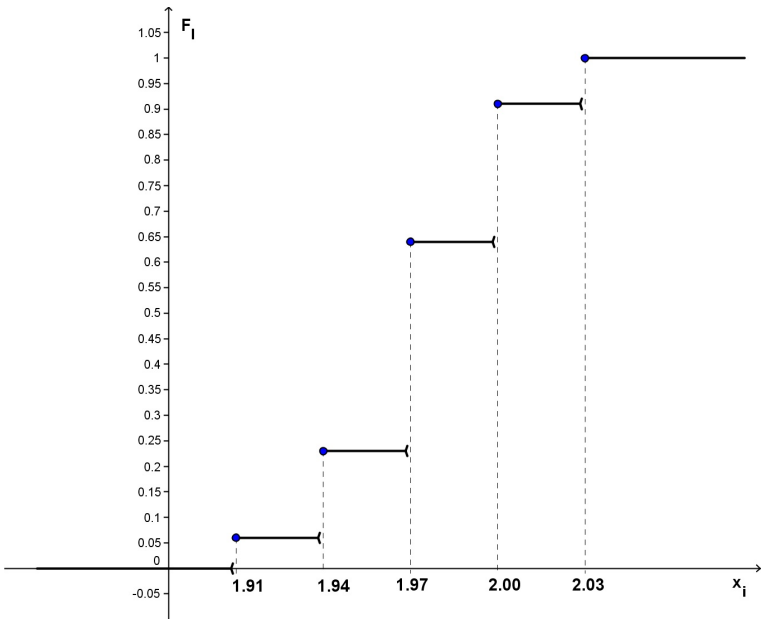

FIGURE 1.5 – Courbe des fréquences cumuleés

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. Pour N l'effectif de la population ou de l'échantillon, on peut le calculer selon l'une des deux régles suivantes :

i) Règle de Sturge : $P = 1 + 3.3 \times \log_{10}(N)$

ii) Règle de Yule : $P = 2.5 \times \sqrt[4]{N}$

Avec $p =$ l'entier le plus proche de P.
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.

Pour la méthode de Sturge $P = 1 + 3.3 \times \log_{10}(20) = 5.293$.
Pour la méthode de Yule $P = 2.5 \times \sqrt[4]{20} = 5.287$,
D'où le nombre de classe est $p = 5$.
Nous avons $x_{min} = 5.11$ et $x_{max} = 6.11$. D'ou $e = 6.11 - 5.11 = 1$ 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.

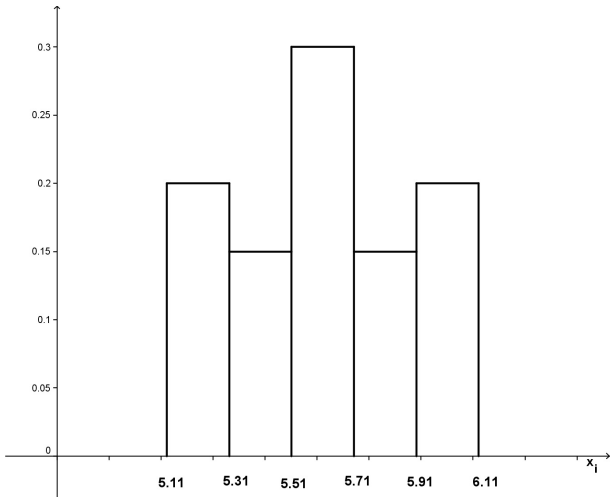

FIGURE 1.6 – Histogramme

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 $c_i = \dfrac{a_i + a_{i+1}}{2}$ le centre de la classe i, $i = 1, \cdots, p$.
Lorsque la borne inférieure de la première (resp. supérieure de la dernière) classe est observée c'est à dire l'intervalle est fermé en a_1 (resp. a_{p+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) (resp. $(c_p,\, f_p)$ et $(c_{p+1}, 0)$) où $c_0 = a_1 - \dfrac{a}{2}$ (resp $c_{p+1} = a_{p+1} + \dfrac{a}{2}$).
Lorsque la borne inférieure de la première (resp. la borne supérieure de la dernière) classe n'est pas observée c'est à dire l'intervalle est ouvert en a_1 (resp. en a_{p+1}) , on complète la courbe en joignant les points $(a_1,\, 0)$ et (c_1, f_1) (resp. $(c_p,\, f_p)$ et $(a_{p+1}, 0)$).

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

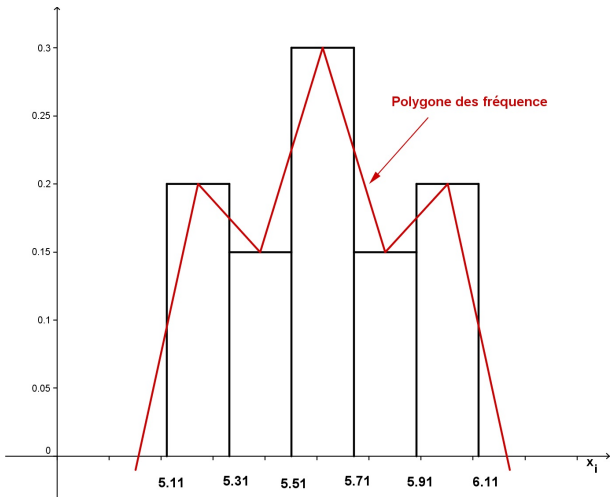

FIGURE 1.7 – Polygône des fréquences

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.

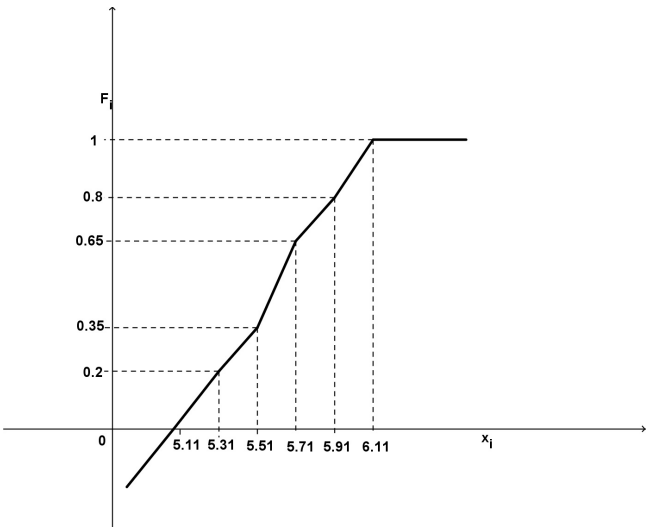

FIGURE 1.8 – Courbe des fréquences cumulées

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

Définition 2.1.1 *:* 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 la fréquence ou l'effectif est le plus élevé.

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[.
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) avant de :

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

le mode m_d (qui appartient à la classe modale) est déterminé par interpolation linéaire. Pour illustrer une telle interpolation, considérons l'exemple suivant : 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.
Nous allons utiliser l'histogramme pour déterminer m_d. En utilisant les triangles d'une part ABC

et CIC_1 et d'autre part ADB et BIC_1 de la figure ci-dessous on a

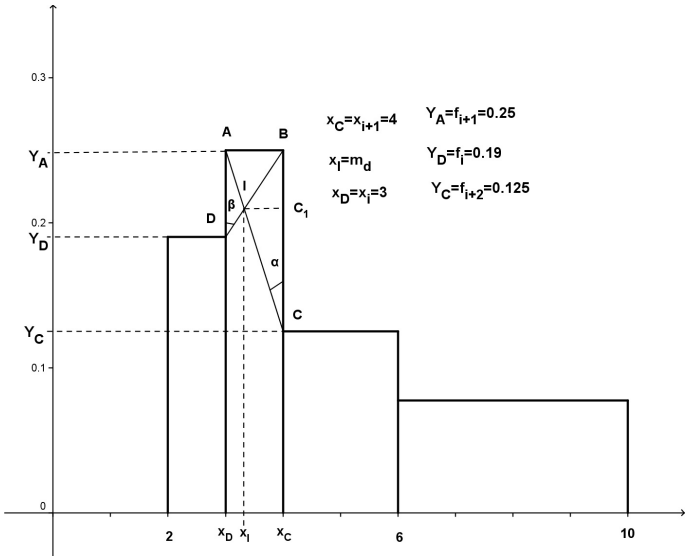

FIGURE 2.1 – Histogramme

$$cotg(\alpha) = \frac{BC}{AB} = \frac{C_1C}{C_1I} \Longrightarrow \frac{Y_A - Y_C}{a} = \frac{Y_I - Y_C}{x_C - x_I}$$

$$cotg(\beta) = \frac{AD}{AB} = \frac{C_1B}{C_1I} \Longrightarrow \frac{Y_A - Y_D}{a} = \frac{Y_A - Y_I}{x_C - x_I}$$

d'où le système $\begin{cases} \dfrac{Y_I - Y_C}{x_C - x_I} = \dfrac{Y_A - Y_c}{a} \\ \dfrac{Y_A - Y_I}{x_C - x_I} = \dfrac{Y_A - Y_D}{a} \end{cases}$

en faisant la somme on obtient

$$\frac{Y_A - Y_C}{x_C - x_I} = \frac{(Y_A - Y_C) + (Y_A - Y_D)}{a}$$

On en déduit

$$\frac{x_C - x_I}{Y_A - Y_C} = \frac{a}{(Y_A - Y_C) + (Y_A - Y_D)} \quad \text{ou encore} \quad x_I = x_C - \frac{a(Y_A - Y_C)}{(Y_A - Y_C) + (Y_A - Y_D)}$$

Où x_{i+1} est la borne supérieure de la classe modale, a l'amplitude commune à toutes les classes, f_{i+1} la fréquence de la classe modale, f_i la fréquence de la classe qui précède la classe modale et f_{i+2} la fréquence de la classe qui suit la classe modale.

$$m_d = x_{i+1} - a \times \frac{(f_{i+1} - f_{i+2})}{(f_{i+1} - f_{i+2}) + (f_{i+1} - f_i)} \text{ ou } m_d = x_{i+1} - a \times \frac{(n_{i+1} - n_{i+2})}{(n_{i+1} - n_{i+2}) + (n_{i+1} - n_i)}$$

Application numérique : $x_{i+1} = 4$, $a = 1$, $f_i = 0.19$, $f_{i+1} = 0.25$ et $f_{i+2} = 0.125$, on a

$$m_d = 4 - 1 \times \frac{(0.25 - 0.125)}{(0.25 - 0.125) + (0.25 - 0.19)} = 3.324$$

2.1.2 La médiane

La médiane est la valeur m de la variable 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 2.1.3 *:*

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 $\dfrac{N}{2} = 15$ et $m = \dfrac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} = \dfrac{x_{(15)} + x_{(16)}}{2} = \dfrac{30+40}{2} = 35$.
$x_{(16)} = 40$ 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	5	8	3	4
effectifs cumulés	4	13	18	26	29	33

On a $N = 33$

donc N est impair d'où $\dfrac{N+1}{2} = 17$ et $m = x_{(17)} = 30$ car le premier effectif cumulé supérieur ou égal à 17 est 18 et $x_{(18)} = 30$.

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

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

La médine m (qui appartient à la classe médiane) est ensuite déterminée à partir d'une interpolation linéaire. 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	

On a $F(4) = 0,44 < 0.5$ et $F(6) = 0.64 > 0.5$, la classe médiane est donc $]4,6]$. Nous utiliserons la courbe des fréquences cumulées pour déterminer m. En considérant les triangles ABD et AIC de la figure ci-dessous, on a

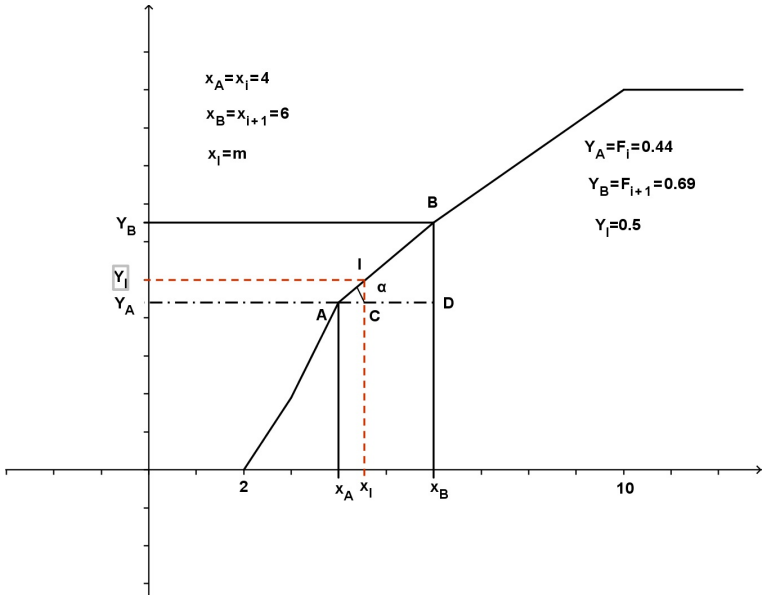

FIGURE 2.2 – Courbe des fréquences cumulées

$$\begin{aligned} tg(\alpha) &= \frac{DB}{AD} = \frac{Y_B - Y_A}{x_B - x_A} = \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i} \\ &= \frac{CI}{AC} = \frac{Y_I - Y_A}{x_I - x_A} = \frac{0,5 - F(x_i)}{m - x_i} \end{aligned}$$

d'où $$m = x_i + (x_{i+1} - x_i)\frac{0,5 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

Application numérique : $x_i = 4$, $x_{i+1} = 6$, $F_i = 0.44$, $F_{i+1} = 0.69$ et

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

2.1.3 Moyennes

2.1.3.1 Moyenne arithmétique

i) 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 2.1.4 *:*

Considérons la distribution de l'exemple 2.1.3 i)

$$\overline{x} = \frac{10 \times 3 + 20 \times 8 + 30 \times 4 + 40 \times 9 + 50 \times 3 + 60 \times 3}{3 + 8 + 4 + 9 + 3 + 3} = \frac{1000}{30} = 33.333$$

ii) 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 $\overline{x} = \sum\limits_i f_i c_i$

Exemple 2.1.5 *:*

Reprenons l'exemple de la distribution des salaires mensuels

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

2.1.3.2 Moyenne quadratique

i) Variable quantitative discrète

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

$$\overline{x}_q = \frac{\sum\limits_i n_i\, x_i^2}{\sum\limits_i n_i} = \frac{\sum\limits_i n_i\, x_i^2}{N} = \sum_i f_i\, x_i^2 \quad (\text{ car } f_i = \frac{n_i}{N})$$

Exemple 2.1.6 *:*

Considérons la distribution de l'exemple 2.1.3 i)

$$\overline{x}_q = \frac{10^2 \times 3 + 20^2 \times 8 + 30^2 \times 4 + 40^2 \times 9 + 50^2 \times 3 + 60^2 \times 3}{3+8+4+9+3+3} = \frac{39800}{30} = 1326.667$$

ii) Variable quantitative continue

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

$$\overline{x}_q = \frac{\sum\limits_i n_i\, c_i^2}{\sum\limits_i n_i} = \frac{\sum\limits_i n_i\, c_i^2}{N} = \sum_i f_i\, c_i^2 \quad (\text{ car } f_i = \frac{n_i}{N})$$

où c_i est le centre de de la classe associée à l'effectif n_i.

Exemple 2.1.7 *:*

Reprenons l'exemple de la distribution des salaires mensuels

$$\overline{x}_q = \frac{15 \times 2.5^2 + 20 \times 3.5^2 + 20 \times 5^2 + 24 \times 8^2}{15+20+20+24} = \frac{2374.75}{79} = 30.060$$

2.1.3.3 Moyenne géométrique

i) Variable quantitative discrète

La moyenne géométrique notée $\overline{x}_G$, d'une variable quantitative discrète est donnée par :

$$\overline{x}_G = \sqrt[N]{\prod_i x_i^{n_i}} \text{ où } N = \sum_i n_i$$

Exemple 2.1.8 *:*

Considérons la distribution de l'exemple 2.1.3 i)

$$\begin{aligned}
\overline{x}_G &= \sqrt[30]{10^3 \times 20^8 \times 30^4 \times 40^9 \times 50^3 \times 60^3} \\
&= \sqrt[33]{10^3 \times (256 \times 10^8) \times (81 \times 10^3) \times (262144 \times 10^9) \times (125 \times 10^3) \times (216 \times 10^3)} \\
&= \sqrt[30]{146767085568000 \times 10^{30}} = 29.663
\end{aligned}$$

ii) Variable quantitative continue

Dans ce cas La moyenne géométrique est donnée par :

$$\overline{x}_G = \sqrt[N]{\prod_i c_i^{n_i}} \text{ où } c_i \text{ est le centre de la classe associée à l'effectif } n_i$$

Exemple 2.1.9 *:*

Reprenons l'exemple de la distribution des salaires mensuels

$$\overline{x}_G = \sqrt[79]{2,5^{15} \times 3,5^{20} \times 5^{20} \times 8^{24}} = 4.6120$$

Remarque : Le logarithme de la moyenne géométrique est égale à la moyenne arithmétique du logaritme de la variable.

$$\ln(\overline{x}_G) = \frac{\sum\limits_i n_i \ln(x_i)}{N} \text{ où } \ln(\overline{x}_G) = \frac{\sum\limits_i n_i \ln(c_i)}{N}$$

2.1.3.4 Moyenne harmonique

i) Variable quantitative discrète

C'est linverse de la moyenne arithmétique des inverses des valeurs de la variable. On la note $\overline{x}_H$,

$$\overline{x}_H = \frac{N}{\sum_i n_i/x_i}$$

ii) Variable quantitative continue

Dans ce cas La moyenne harmonique est donnée par :

$$\overline{x}_H = \frac{N}{\sum_i n_i/c_i}$$

Remarque : $\overline{x}_H \leq \overline{x}_G \leq \overline{x} \leq \overline{x}_q$

2.2 Les mesures de dispersion

Les indicateurs de dispersion sont nombreux, les plus courants sont :
L'étendue, l'écart interquartile, la variance, l'ecart-type 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 $\quad \overline{x} = \frac{4\times 10 + 1\times 20 + 4\times 30}{9} = \frac{180}{9} = 20$

b) 20, 22, 21, 20, 20, 19, 18, 20, 20 $\quad \overline{x} = \frac{18+19+5\times 20+21+22}{9} = \frac{180}{9} = 20$

c) 1, 4, 6, 8, 20, 32, 34, 36, 39 $\quad \overline{x} = \frac{1+4+6+8+20+32+34+36+39}{9} = \frac{180}{9} = 20$

d) 10, 12, 14, 16, 20, 24, 26, 28, 30 $\quad \overline{x} = \frac{10+12+14+16+20+24+26+28+30}{9} = \frac{180}{9} = 20$

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 :
$Etendue(a) = 30 - 10 = 20$, $Etendue(b) = 22 - 18 = 4$, $Etendue(c) = 39 - 1 = 38$ et
$Etendue(d) = 30 - 10 = 20$.
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 la borne supérieure de la dernière classe et la borne inférieure de la première classe.

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

où x_{max} (resp. x_{min}) est la borne supérieure (resp. inférieure) de la dernière (resp. première) classe.

2.2.2 Les quartiles

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 $2^{ème}$ quartile, le $5^{ème}$ décile et le $50^{ème}$ centile.

2.2.2.1 Variable quantitative discrète

On cosidère une série statistique dont les valeurs du caractère étudié, ont été rangés dans un ordre croissant :

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_{n-1} \leq x_n$$

La médiane m_e sépare la série en deux séries de même effectif. La série inférieure dont les valeurs du caractère sont inférieures ou égale à la médiane m_e, et la série supérieure dont les valeurs du caractère sont supérieures ou égale à la médiane m_e.
On appelle premier (resp. troisième) quartile, la médiane de la série inférieure (resp. supérieure) on le note Q_1 (resp. Q_3).

Exemple 2.2.2 *:*

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 $N = 30$ et $m = 35$

x_i	10	20	30
n_i	3	8	4
effectifs cumulés	3	11	15

série inférieure avec $N_1 = 15$

x_i	40	50	60
n_i	9	3	3
effectifs cumulés	9	12	15

série supérieure avec $N_1 = 15$

donc N_1 est impair d'où $\dfrac{N_1+1}{2} = 8 \Longrightarrow Q_1 = x_{(\frac{N_1+1}{2})} = x_{(8)} = 20$ et $Q_3 = x_{(\frac{N_1+1}{2})} = x_{(8)} = 40$.

ii) Considérons la distribution suivante

x_i	10	20	30	40	50	60
n_i	4	9	5	8	3	4
effectifs cumulés	4	13	18	26	29	33

On a $N = 33$ et $m = 30$.

x_i	10	20	30
n_i	4	9	3
effectifs cumulés	4	13	16

série inférieure avec $N_1 = 16$

x_i	30	40	50	60
n_i	1	8	3	4
effectifs cumulés	1	9	12	16

série supérieure avec $N_1 = 16$

donc N_1 est pair d'où $\frac{N_1}{2} = 8 \Longrightarrow Q_1 = \frac{x_{(\frac{N_1}{2})} + x_{(\frac{N_1}{2}+1)}}{2} = \frac{x_{(8)} + x_{(9)}}{2} = 20$ et
$Q_3 = \frac{x_{(\frac{N_1}{2})} + x_{(\frac{N_1}{2}+1)}}{2} = \frac{x_{(8)} + x_{(9)}}{2} = \frac{40+40}{2} = 40.$

2.2.2.2 Variable quantitative continue

Des techniques similaires à celles utilisées pour déterminer la médiane dans le cas continue permettent de déterminer ces indicateurs.
Pour le premier quartile

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

Pour le troisième quartile

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

Exemple 2.2.3 *:*
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	

$0.19 < F(Q_1) = 0.25 \leq 0.44 \Longrightarrow 3 < Q_1 \leq 4$, d'où $Q_1 = 3 + (4-3) \times \frac{0,25 - 0,19}{0,44 - 0,19} = 3,24$
$0.69 < F(Q_3) = 0.75 \leq 1 \Longrightarrow 6 < Q_3 \leq 10$, d'où $Q_3 = 6 + (10-6) \times \frac{0,75 - 0,69}{1 - 0,69} = 6,19.$

2.2.2.3 L'écart interquartile

Q_1 étant le premier quartile et Q_3 le troisième quartile, l'écart interquartile est la différence entre le troisième et le premier quartile, il est noté $\mathbf{R}(Q) = Q_3 - Q_1$.
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.
L'écart interquartile $\mathbf{R}(Q)$ est la largeur de l'intervalle interquartile. C'est une mesure de longueur de cet intervalle et donc une mesure de dispersion des données autours de la médiane.

- Plus il est grand, plus les données sont dispersées autours de la médiane.
- Plus il est petit, plus les données sont proches de la médiane.

Exemple 2.2.4 *:*
Reprenons l'exemple de la distribution des salaires mensuels.
L'intervalle interquartile est $[3,24,\ 6,19]$ et l'écart interquartile est $\mathbf{R}(Q) = 6,19 - 3,24 = 2,85$.

2.2.3 Diagramme en boîte

Ce diagramme est aussi appelé boîte à moustaches. Il utilise la valeur du 1^{er} quartile Q_1 (qui correspond à 25% des effectifs), la valeur du $2^{ème}$ quartile $Q_2 = m$ (la médiane qui correspond à 50% des effectifs), la valeur du $3^{ème}$ quartile Q_3 (qui correspond à 75% des effectifs), l'écart interquartile $\mathbf{R}(Q)$ et les valeurs minimum et maximum de la série.
On représente sur un axe gradué (horizontal ou vertical) les différentes valeurs de la série Q_1, Q_2, Q_3, x_{min}, x_{max} ainsi que $Q_1 - 1.5 \times \mathbf{R}(Q)$ et $Q_3 + 1.5 \times \mathbf{R}(Q)$.
Le diagramme est formé d'un rectangle ayant pour extrémité inférieure le 1^{er} quartile et pour extrémité supérieure le $3^{ème}$ quartile. A l'interieur de ce rectangle, on trace un segment représentant la médiane.
A gauche et à droite de ce rectangle, on trace deux segments appelé "moustaches" inférieure et supérieure qui ont pour extrémité respectivement $Q_1 - 1.5 \times \mathbf{R}(Q)$ et $Q_3 + 1.5 \times \mathbf{R}(Q)$.

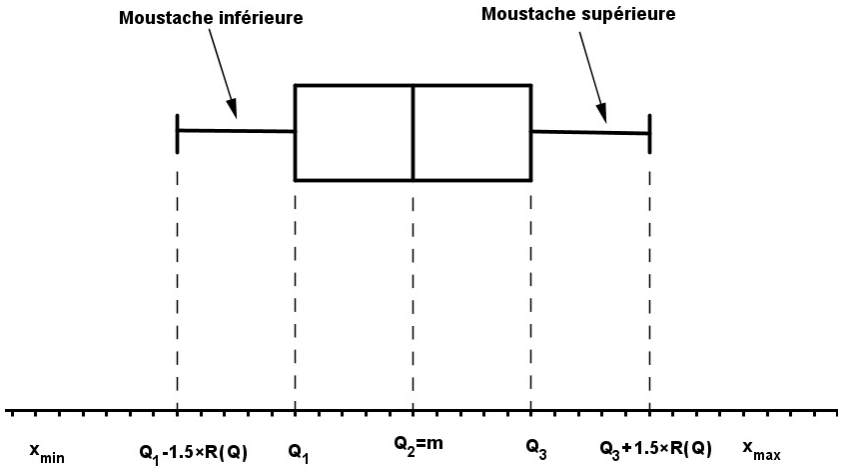

FIGURE 2.3 – Boîte à moustaches

La boîte a pour largeur l'écart interquartile $(Q_3 - Q_1)$. les moustaches sont basées généralement sur 1.5 fois la largeur de la boîte. Dans ce cas, une valeur est atypique ou aberrante si elle dépasse de 1.5 fois l'écart interquartile à gauche du 1^{er} quartile ou à droite du $3^{ème}$ quartile.
La boîte à moustaches permet de répondre à certaines questions :

- Existe-t-il des observations atypiques ? en les repérant et les identifiant
- La distribution est-elle symétrique ? en repérant la position de la médiane dans la boîte.
- La partie centrale (50% des effectifs) est-elle plus ou moins concentrée ou étalée par rapport au reste de la distribution ?
- Comparaisons de distributions selon des groupes ? Pour comparer les distributions d'une même variable selon les groupes, on juxtapose sur le même graphique les boîtes à moustaches définies respectivement pour les groupes en utilisant la même échelle.

Exemple 2.2.5 *:*
Deux groupes de S3 Statistique comparent leurs résultats du Contrôle final et déclarent : "nos classes ont le même profil puisque dans les deux cas la médiane et le mode des résultats est 10".

Qu'en pensez-vous ?

notes	5	6	7	8	9	10	11	12	13	14	15	16	17
groupe 1	4	4	3	3	3	4	3	2	2	3	2	2	1
groupe 2	1	3	4	4	5	7	4	3	1	2	1	0	2

Vérifier que les deux médianes valent 10 et déterminer les quartiles de chaque série. Tracer côte à côte les diagrammes en boites de ces deux séries.
Les effectifs cumulés des deux groupes est :

notes	5	6	7	8	9	10	11	12	13	14	15	16	17
groupe 1	4	8	11	14	17	21	24	26	28	31	33	35	36
groupe 2	1	4	8	12	17	24	28	31	32	34	35	35	37

$N_1 = 36$ est pair d'où $\dfrac{N_1}{2} = 18 \Longrightarrow m_1 = \dfrac{x_{(\frac{N_1}{2})} + x_{(\frac{N_1}{2}+1)}}{2} = \dfrac{x_{(18)} + x_{(19)}}{2} = \dfrac{10+10}{2} = 10.$
$N_2 = 37$ est impair d'où $\dfrac{N_2+1}{2} = 19 \Longrightarrow m_2 = x_{(\frac{N_2+1}{2})} = x_{(19)} = 10.$
Les séries inférieures et supérieurs du groupe 1 et 2 sont :

notes	5	6	7	8	9	10
groupe 1	4	4	3	3	3	1
groupe 2	1	3	4	4	5	1

série inférieure avec $N_{i1} = N_{i2} = 18$

notes	10	11	12	13	14	15	16	17
groupe 1	3	3	2	2	3	2	2	1
groupe 2	5	4	3	1	2	1	0	2

série supérieure avec $N_{s1} = N_{s2} = 18$

Les effectifs des séries inférieures et supérieurs du groupe 1 et 2 sont :

notes	5	6	7	8	9	10
groupe 1	4	8	11	14	17	18
groupe 2	1	4	8	12	17	18

série inférieure avec $N_{i1} = N_{i2} = 18$

notes	10	11	12	13	14	15	16	17
groupe 1	3	6	8	10	13	15	17	18
groupe 2	5	9	12	13	15	16	16	18

série supérieure avec $N_{s1} = N_{s2} = 18$

On a $N_{i1} = N_{i2} = 18$ est pair d'où :
$Q_{11} = \dfrac{x_{(\frac{N_{i1}}{2})} + x_{(\frac{N_{i1}}{2}+1)}}{2} = \dfrac{x_{(9)} + x_{(10)}}{2} = 7$ et $Q_{12} = \dfrac{x_{(\frac{N_{i2}}{2})} + x_{(\frac{N_{i2}}{2}+1)}}{2} = \dfrac{x_{(9)} + x_{(10)}}{2} = 8.$
On a $N_{s1} = N_{s2} = 18$ est pair d'où :
$Q_{31} = \dfrac{x_{(\frac{N_{s1}}{2})} + x_{(\frac{N_{s1}}{2}+1)}}{2} = \dfrac{x_{(9)} + x_{(10)}}{2} = 13$ et $Q_{32} = \dfrac{x_{(\frac{N_{s2}}{2})} + x_{(\frac{N_{s2}}{2}+1)}}{2} = \dfrac{x_{(9)} + x_{(10)}}{2} = 11.5.$
L'écart interquartile des deux groupes est : $\mathbf{R}(Q1) = 13 - 7 = 6$ et $\mathbf{R}(Q2) = 11.5 - 8 = 3.5.$
$$\Longrightarrow \begin{cases} Q_{11} - 1.5 \times \mathbf{R}(Q1) = -2 & Q_{31} + 1.5 \times \mathbf{R}(Q1) = 22 \\ Q_{12} - 1.5 \times \mathbf{R}(Q2) = 2.75 & Q_{32} + 1.5 \times \mathbf{R}(Q2) = 16.75 \end{cases}$$

Le graphique ci-dessous met bien en évidence que l'écart interquartile est plus resserré pour le groupe 2 que le groupe 1 donc les élèves du groupe 2 ont globalement un niveau plus homogène que ceux de du groupe 1. On peut remarquer que 17 est une valeur atypique pour le groupe 2 tandis que le groupe 1 n'a pas de valeur atypique. La distribution du groupe 1 est symetrique car la boîte est symetrique par rapport au segment de la médiane tandis que celle du groupe 2 est asymetrique à gauche.

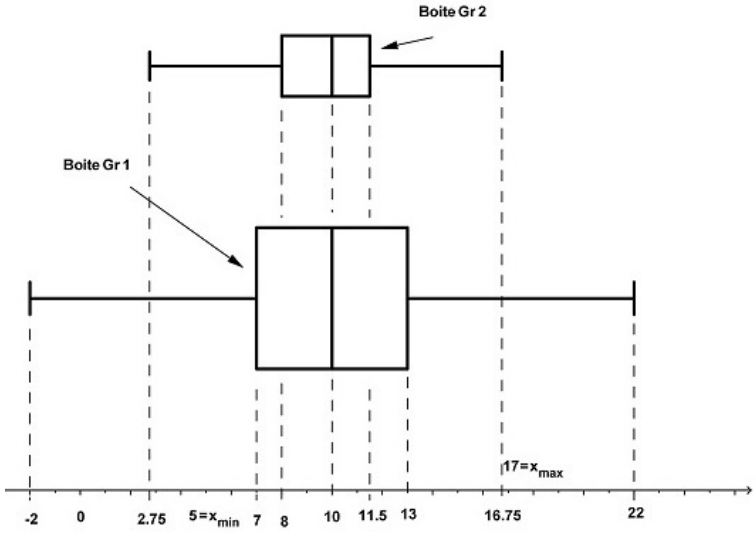

FIGURE 2.4 – Boîte à moustaches des Gr 1 et 2

2.2.4 Diagramme tige et feuille

Un diagramme " tige et feuille " est une autre façon de résumer et représenter un ensemble de données de la distribution d'une variable quantitative. C'est un diagramme plus instructif pour les bases de données relativement petites (moins de 100 unités). Il se situe à mi chemin entre le tableau de distribution et le graphique. Comment construire un diagramme " tige et feuille " ?

• Séparer chaque nombre en une tige qui contient tous les chiffres sauf le dernier et une feuille, soit le dernier chiffre. Les tiges ont autant de chiffres que nécessaire, alors que la feuille n'a qu'un seul chiffre.

• On place les tiges sur une colonne verticale avec la plus petite tige en haut.

• On écrit chaque feuille à droite de sa tige en ordre croissant.

Notons qu'une valeur est répétée autant de fois qu'elle apparaît.

Les avantages d'une telle présentation sont multiples :

• Toutes les valeurs y sont nommées et ordonnées

• Ce tracé ressemble quand on le tourne à un diagramme en bâtons.

• On peut y ajouter l'effectif de chaque tige.

• On peut y lire facilement le nombre de données, la valeur la plus grande, la plus petite, la plus fréquente ainsi que les éventuelles valeurs aberrantes.

• On peut repérer facilement la médiane, les quartiles, les déciles.

• On peut remarquer la symétrie ou l'asymétrie (lorsque sa forme générale est désaxée).

Exemple 2.2.6 *:*

On considère une série de taux d'hémoglobine dans le sang (en $g.l^{-1}$) mesuré chez des adultes présumés en bonne santé. La série ordonnée est :

105 110 112 112 118 119 120 120 125 125 126 127 128 130 132 133 134 135 138 138 138 138 141 142 144 145 146 148 148 148 149 150 150 150 151 151 153 153 153 154 154 154 155 156 156 158 160 160 160 163 164 164 165 166 166 168 168 170 172 172 176 179. Un tracé en tiges et feuilles donne :

Tige	Feuille															Effectifs
10	5															1
11	0	2	2	8	9											5
12	0	0	5	5	6	7	8									7
13	0	2	3	4	5	8	8	8	8							9
14	1	2	4	5	6	8	8	8	9							9
15	0	0	0	1	1	3	3	3	4	4	4	5	6	6	8	15
16	0	0	0	3	4	4	5	6	6	8	8					11
17	0	2	2	6	9											5

On peut lire ainsi que la valeur 105 est la plus petite valeur qui semble être une valeur aberrante, que 179 la plus grande valeur, que 120 figure 2 fois dans la série, 138 figure 4 fois. Pour calculer la médiane , on a $N = 62$ pair et $\frac{N}{2} = 31 \Longrightarrow m = \frac{x_{(31)} + x_{(32)}}{2} = \frac{149 + 150}{2} = 149.5$, pour calculer le 1^{er} quartile, on a $\frac{N}{2} = 31$ impair et $\frac{\frac{N}{2} + 1}{2} = 16 \Longrightarrow Q_1 = x_{(16)} = 133$ et pour calculer le $3^{ème}$ quartile, on a $\frac{N}{2} = 31$ impair et $\frac{\frac{3N}{2} + 1}{2} = 47 \Longrightarrow Q_3 = x_{(47)} = 160$.

Un diagramme dos à dos de tige et feuille peut être employé pour comparer deux bases de données. Ci-dessous, nous représentons les notes sur 100 de deux groupes du cours de statistique d'un examen en utilisant le diagramme dos à dos de tige et feuille :

Groupe A														Groupe B										
Effectifs	Feuille												Tige	Feuille										Effectifs
0													0	5										1
2											3	1	2	4	5	7								3
4									4	4	3	3	3	1	2	2	8	8	9					6
5								9	9	6	6	4	4	3	3	3	4	7	7	7				7
10			7	5	5	4	4	4	4	2	2	1	5	4	4	4	6	6	8	8	8	9		9
12	9	9	8	7	7	7	3	3	2	1	1	1	6	1	2	4	4	5	5	9				7
6							9	8	7	5	5	2	7	3	3	4	6	6	6					6
6							6	6	6	3	1	1	8	2	5	9								3
3										4	3	2	9	1										1

2.2.5 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. L'écart-type permet d'avoir une idée de la façon dont les valeurs de la série s'écatent par rapport à la moyenne, c'est donc une mesure de dispersion. Un écart-type faible correspond à une série concentrée autour de la moyenne.

2.2.5.1 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 2.2.7 *:*
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 : $\sum_i n_i(x_i - \overline{x})^2 = \sum_i n_i x_i^2 - N\overline{x}^2 \Longrightarrow V(x) = \frac{1}{N}\left(\sum_i n_i x_i^2\right) - \overline{x}^2$

2.2.5.2 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 2.2.8 *:*
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 $\overline{x} = 5.05$

$$\begin{aligned} V(x) &= \frac{15(2.5-5.05)^2 + 20(3.5-5.05)^2 + 20(5-5.05)^2 + 24(8-5.05)^2}{79} \\ &= \frac{354.497}{79} = 4.487 \\ \sigma(x) &= \sqrt{4.487} = 2.118 \end{aligned}$$

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

2.2.6 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 2.2.9 *:*
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.

2.2.7 Moments

Définition 2.2.1 *:* Le moment d'ordre r d'une variable statistique est la quantité

$$m_r = \frac{1}{N}\sum_i n_i\, x_i^r \;\text{ ou } m_r = \frac{1}{N}\sum_i n_i\, c_i^r \text{ où } N = \sum_i n_i$$

Pour $r = 0$, $m_0 = 1$.
Pour $r = 1$, $m_1 = \overline{x}$ la moyenne arithmétique.

Définition 2.2.2 *:* Le moment centré d'ordre r d'une variable est la quantité

$$\mu_r = \frac{1}{N}\sum_i n_i\,(x_i - \overline{x})^r \;\text{ ou } \mu_r = \frac{1}{N}\sum_i n_i\,(c_i - \overline{x})^r \text{ où } N = \sum_i n_i$$

Pour $r = 0$, $\mu_0 = 1$.
Pour $r = 1$, $\mu_1 = 0$
Pour $r = 2$, $\mu_2 = V(x)$ la variance.

2.2.8 Changement d'origine et d'unité

2.2.8.1 Changement d'origine et d'unité

Définition 2.2.3 *:*

• On appelle changement d'origine l'opération consistant à ajouter la même quantité $b \in \mathbb{R}$ à toutes les observations : $y_i = x_i + b,\ i = 1, \cdots, n$.

• On appelle changement d'unité l'opération consistant à multiplier par la même quantité $a \in \mathbb{R}$ toutes les observations : $y_i = a \times x_i,\ i = 1, \cdots, n$.

• On appelle changement d'origine et d'unité l'opération consistant à multiplier toutes les observations par la même quantité $a \in \mathbb{R}$ puis à ajouter la même quantité $b \in \mathbb{R}$ à toutes les observations : $y_i = a \times x_i + b,\ i = 1, \cdots, n$.

Théorème 2.2.1 *:* Si on effectue un changement d'origine et d'unité sur une variable X, alors

• Sa moyenne est affectée du même changement d'origine et d'unité, $\overline{y} = a\,\overline{x} + b$

• Sa variance et son écart-type sont affectés par le changement d'unité et pas par le changement d'origine, $V_y = a^2\,V_x$ et $\sigma_y \sqrt{V_y} = |a|\,\sigma_x$

Preuve : Si $y_i = a \times x_i + b$, alors

$$\begin{aligned}
\overline{y} &= \frac{1}{n}\sum_{i=1}^{n}(a \times x_i + b) = a \times (\frac{1}{n}\sum_{i=1}^{n} x_i) + b = a\,\overline{x} + b \\
V_y &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(a \times x_i + b - a\overline{x} - b)^2 = a^2\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2\,V_x \\
\sigma_y &= \sqrt{V_y} = \sqrt{a^2\,V_x} = |a|\,\sigma_x
\end{aligned}$$

Remarque :

• Les paramètres de position (mode, médiane et moyenne) sont tous affectés par un changement d'origine et d'unité.

• Les paramètres de dispersion sont tous affectés par un changement d'unité mais pas par un changement d'origine (sauf le coefficient de variation).

2.2.8.2 Centrer et réduire une variable

Centrer et réduire une variable statistique quantitative X consiste la remplacer par la variable : $\dfrac{X - \overline{X}}{\sigma(X)}$.

$X - \overline{X}$ pour la centrer (moyenne 0). La variable : $\dfrac{X - \overline{X}}{\sigma(X)}$ a pour moyenne arithmétique 0 elle est centré.

Diviser par l'écart-type $\sigma(X)$ pour la réduire (écart-type = 1). La variable $\dfrac{X - \overline{X}}{\sigma(X)}$ a pour variance et écart-type 1 elle est réduite.

2.3 Parmètre de forme

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.

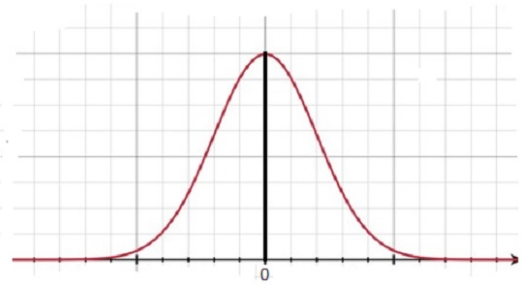

FIGURE 2.5 – Cloche

A la différence de la médiane et du mode, 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 et le mode. 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 lorsque la distribution ne présente qu'un seul mode.

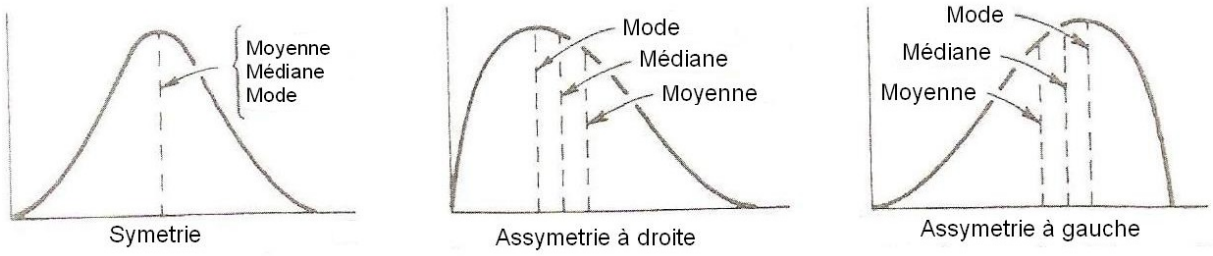

FIGURE 2.6 – symetrie et asymetrie

2.3.2 Coefficient d'asymétrie

le coefficient d'asymétrie a pour rôle de fournir une mesure de dissymétrie d'une distribution.

2.3.2.1 Coefficient de d'asymétrie de Pearson

Le premier coefficient d'asymétrie de Pearson est basé sur une comparaison de la moyenne et de la médiane, et est normalisé par l'écart-type. Il est calculé à partir de la formule suivante :

$$A_{P1} = 3 \times \frac{\overline{x} - m}{\sigma} \quad \text{où } \overline{x} \text{ est la moyenne, } m \text{ la médiane et } \sigma \text{ l'écart-type.}$$

Lorsque la distribution statistique est unimodale, on utilise le second coefficient de Pearson basé sur une comparaison de la moyenne et du mode, et est normalisé par l'écart-type. Il est calculé à partir de la formule suivante :

$$A_{P2} = \frac{\overline{x} - m_d}{\sigma} \quad \text{où } \overline{x} \text{ est la moyenne, } m_d \text{ le mode et } \sigma \text{ l'écart-type.}$$

2.3.2.2 Coefficient de d'asymétrie de Yule

Le coefficient d'asymétrie de Yule est basé sur les positions des trois quartile et est normalisé par l'écart interquartile. Il est calculée à partir de la formule suivante :

$$A_Y = \frac{Q_1 + Q_3 - 2 \times Q_2}{\mathbf{R}(Q)} \quad \text{où } Q_1, Q_2, Q_3 \text{ les 3 quartiles , et } \mathbf{R}(Q) \text{ l'écart interquartile.}$$

2.3.2.3 Coefficient de d'asymétrie de Fisher

Le coefficient d'asymétrie de Fisher est basé sur le moment d'ordre 3 et est normalisé par le cube de l'écart-type. Il est calculée à partir de la formule suivante :

$$A_F = \frac{\mu_3}{\sigma^3} \quad \text{où } \mu_3 \text{ le moment centré d'ordre 3 , et } \sigma \text{ l'écart-type.}$$

Tous les coefficients d'asymetrie ont les mêmes propriétés.

• Si la distrbution est symétrique, le coefficient est nul. On admettra que si le coefficient de Fisher $A_F \in]-0.1, 0.1[$, la distribution est symétrique.

• Si la distribution est asymétrique à droite (resp. à gauche) c'est à dire la courbe est étalée à droite (resp. à gauche), le coeficient est positif (resp. négatif).

Remarque : Les paramètres d'asymétrie ne sont pas affectés par un changement d'unité ou d'origine.

Exemple 2.3.1 *:*

On considère la série statistique suivante (masse en grammes des oeufs de poule d'un élevage).

masse : x_i	40	45	50	55	60	65	70	75	80	85	90
Effectif : n_i	16	20	75	141	270	210	165	63	21	12	7

$\overline{x}$	V	σ	μ_3	$m = Q_2$	m_d	Q_1	Q_3	$\mathbf{R}(Q)$	A_{P1}	A_{P2}	A_Y	A_F
62.5	73.8	8.59	91.125	60	60	55	70	15	0.87	0.29	0.33	0.14

La distrbution des masses est asymetrie à droite car les coefficients d'asymetrie sont positifs.

2.3.3 Le Coefficient d'aplatissement

Le coefficient d'aplatissement mesure le degré d'aplatissement d'une distribution. On l'obtient à partir du moment centré d'ordre 4.

- Coefficient d'aplatissement de Pearson

$$\beta_2 = \frac{\mu_4}{V(x)^2} \quad \text{où } V(x) \text{ est la variance et } \mu_4 \text{ le moment centré d'ordre 4}$$

- Coefficient d'aplatissement de Ficher

$$F_2 = \beta_2 - 3 \quad \text{où } \beta_2 \text{ est le coefficient d'aplatissement de Pearson}$$

3 est le degré d'aplatissement d'une loi gaussienne centrée réduite.
Si $F_2 = 0$, le polygone statistique de la variable centrée réduite $\frac{X - \overline{X}}{\sigma}$ à le même aplatissement qu'une courbe en cloche, on dit que la variable est mesokurtique.
Si $F_2 > 0$, le polygone statistique de la variable centrée réduite est moins aplati qu'une courbe en cloche, la concentration des valeurs de la série autour de la moyenne est forte, on dit que la variable est leptokurtique.
Si $F_2 < 0$, le polygone statistique de la variable centrée réduite est plus aplati qu'une courbe en cloche, la concentration des valeurs autour de la moyenne est faible, on dit que la variable est platykurtique.

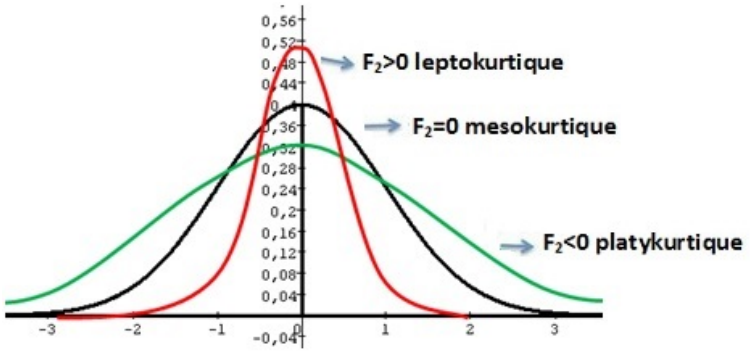

FIGURE 2.7 – Applatissement

Exemple 2.3.2 *:*
Reprenons la distribution des masse des oeufs de poule de l'exemple 2.3.1.
$\mu_4 = 17523.91$, $V(x) = 73.8$, $\beta_2 = 3.22$ et $F_2 = 0.22 > 0 \Longrightarrow$ la variable est leptokurtique et le polygone statistique de la variable centrée réduite est moins applati qu'une courbe en cloche, la concentration des valeurs de la série autour de la moyenne est forte.

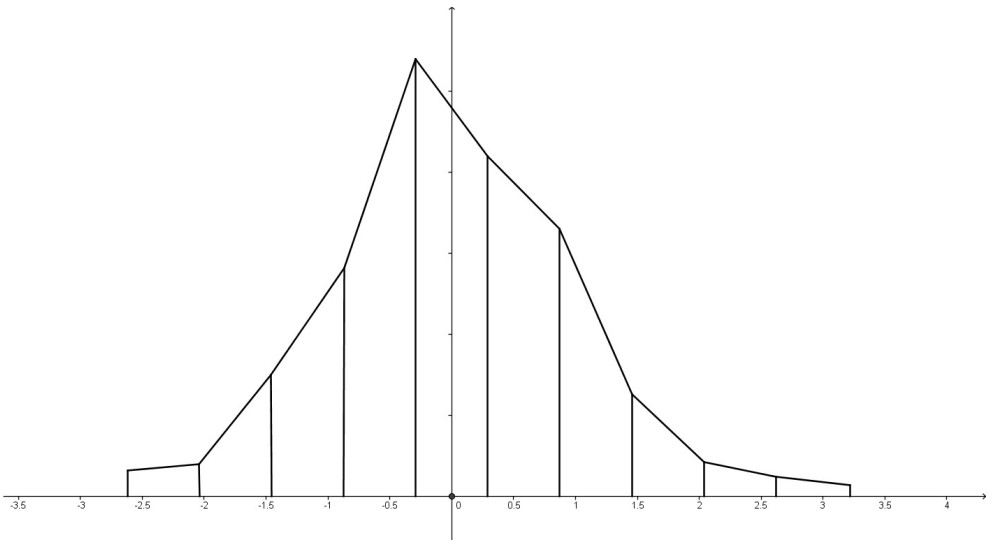

FIGURE 2.8 – Polygône des fréquences de la variable centrée réduite

2.4 Applications : Le théorème de Tchebychev

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 Tchebychev permet d'évaluer le pourcentage des données qui se trouvent à k écart-types de la moyenne c'est 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.

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. x_k et y_k $k = 1, \cdots, n$ sont des valeurs observées.
On représente une distribution statistique à deux caractères quantitatifs par l'ensemble des points A_k, 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, 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).

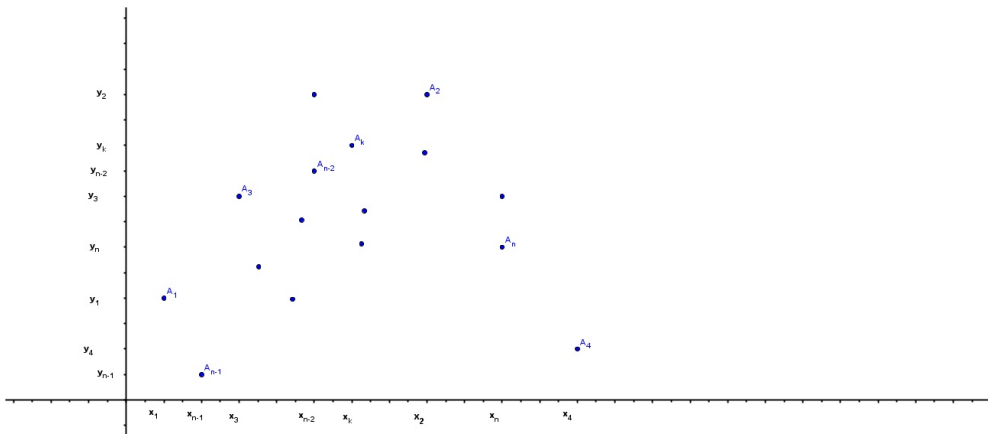

FIGURE 3.1 – Nuage de points

3.2 Ajustement linéaire

L'objectif est de mettre en évidence l'existence d'une relation entre deux variables quantitatives (continues ou discrètes). On cherche un modèle de la forme : $Y = aX + b + \varepsilon$ où :

- Y est la variable dépendante.
- X est la variable explicative (indépendante).
- ε est l'erreur introduite par le modèle (variable centrée).
- a et b les paramètres du modèle avec a la pente de la droite d'ajustement et b l'ordonné à l'origine.
- $y_k^* = a\,x_k + b\ k = 1, \cdots, n$ les valeurs ajustées.
- $e_k = y_k - y_k^*\ k = 1, \cdots, n$ les résidus.

3.2.1 Covariance et coefficient de corrélation

La covariance des variables X et Y s'écrit :

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

avec $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$ et $\overline{y} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} y_k$.
La covariance dépend des unités de mesures dans lesquelles sont exprimées les variables. De même, on définit le coefficient de corrélation :

$$\rho_{xy} = \frac{Cov(x,\, y)}{\sigma(x)\,\sigma(y)} \text{ avec } \sigma(x) \text{ et } \sigma(y) \text{ l'écart-type des variables } X \text{ et } Y$$

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.

Remarque Si $\rho_{xy} > 0$, les points sont alignés le long d'une droite croissante. Si $\rho_{xy} < 0$, les points sont alignés le long d'une droite décroissante. Si $\rho_{xy} = 0$ ou proche de zéro, il n'y a pas de liaison linéaire. On peut cependant avoir une liaison non linéaire avec un coefficient de corrélation nul.

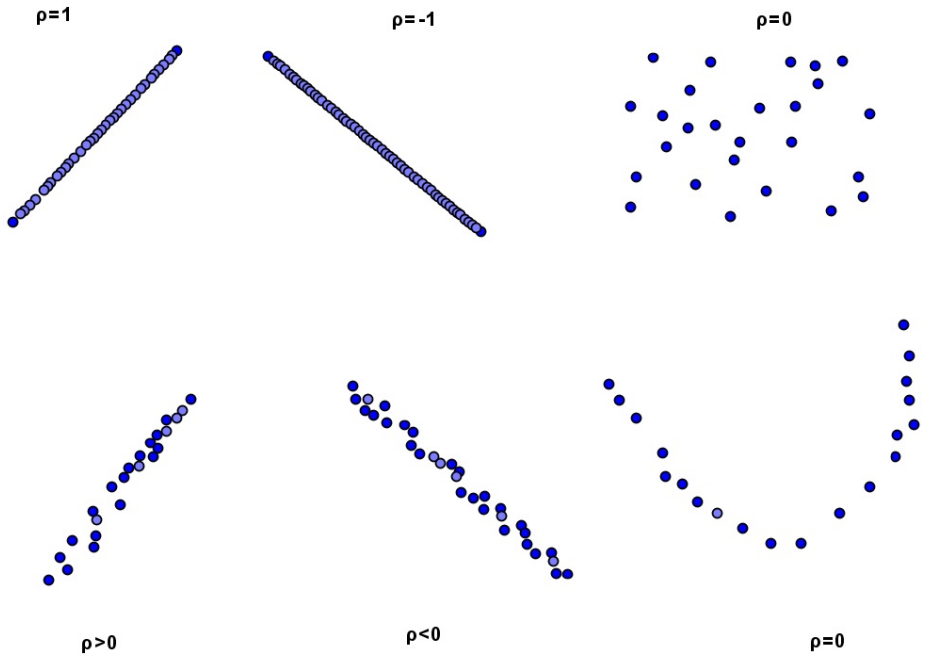

FIGURE 3.2 – Exemples de nuages de points et coefficients de corrélation

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

$$\overline{x} = \frac{10.3}{6} = 1.717 \qquad \overline{y} = \frac{340}{6} = 56.667$$

$$V(x) = \frac{20.75}{6} - 1.717^2 = 0.51 \qquad V(y) = \frac{23050}{6} - 56.667^2 = 630.52$$

$$Cov(x,\, y) = \frac{689.50}{6} - 1.717 \times 56.667 = 17.62 \qquad \rho_{xy} = \frac{17.62}{0.714 \times 25.11} = 0.98$$

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 la plus petite 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

$$\left\{ \begin{array}{rcl} \hat{a} & = & \dfrac{Cov(x,\, y)}{V(x)} \\ \hat{b} & = & \overline{y} - a\,\overline{x} \end{array} \right.$$

La droite $y = \hat{a}\,x + \hat{b}$ est appelée droite de régression linéaire.
Prueve : En annulant les dérivées partielles par rapport à a et b de ϕ, on obtient,

$$\left\{ \begin{array}{rcl} \dfrac{\partial \phi}{\partial a}(a,\, b) & = & -2 \displaystyle\sum_{k=1}^{n} (y_k - a\,x_k - b)\,x_k = 0 \\ \dfrac{\partial \phi}{\partial b}(a,\, b) & = & -2 \displaystyle\sum_{k=1}^{n} (y_k - a\,x_k - b) = 0 \end{array} \right. \Longrightarrow \left\{ \begin{array}{rcl} 0 & = & \displaystyle\sum_{k=1}^{n} (y_k x_k - a\,x_k^2 - b\,x_k) \\ n\,b & = & \displaystyle\sum_{k=1}^{n} (y_k - a\,x_k) = n\,(\overline{y} - a\,\overline{x}) \end{array} \right.$$

$$\Longrightarrow \left\{ \begin{array}{rcl} 0 & = & \displaystyle\sum_{k=1}^{n} y_k x_k - n\,\overline{x}\,\overline{y} - a\,(\sum_{k=1}^{n} x_k^2 - n\overline{x}^2) = n\,Cov(x,\, y) - n\,a\,V(x) \\ b & = & \overline{y} - a\,\overline{x} \end{array} \right. \Longrightarrow \left\{ \begin{array}{rcl} \hat{a} & = & \dfrac{Cov(x,\, y)}{V(x)} \\ \hat{b} & = & \overline{y} - \hat{a}\,\overline{x} \end{array} \right.$$

montrons que le point critique obtenu est un minimum. Calculons les dérivées partielles seconde

$r = \dfrac{\partial^2 \phi}{\partial a^2}(\hat{a},\, \hat{b}) = 2 \sum_{k=1}^{n} x_k^2,\ s = \dfrac{\partial \phi}{\partial a\, \partial b}(\hat{a},\, \hat{b}) = 2 \sum_{k=1}^{n} x_k = 2\, n\, \overline{x},\ t = \dfrac{\partial^2 \phi}{\partial b^2}(\hat{a},\, \hat{b}) = 2\, n$

$s^2 - r\, t = 4\, n^2\, \overline{x}^2 - 4\, n \sum_{k=1}^{n} x_k^2 = -4\, n\, V(x) < 0,\ r > 0 \Longrightarrow \phi$ admet un minimum en $(\hat{a},\, \hat{b})$.

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

3.2.3 Résidus et valeurs ajustées

Les valeurs ajustées sont : $y_k^* = \hat{a}\, x_k + \hat{b}$, $k = 1, \cdots, n$. Ils sont les "prédictions" des y_k réalisées au moyen de la variable X et de la droite de régression de y en x.

Les résidus sont les différences entre les valeurs observées et les valeurs ajustées : $e_k = y_k - y_k^*$, $k = 1, \cdots, n$. Les résidus représentent la partie inexpliquée de la variable Y par la régression. Ils sont de moyenne nulle. En effet, $\dfrac{1}{n} \sum_{k=1}^{n} e_k = \dfrac{1}{n} \sum_{k=1}^{n} (y_k^* - y_k) = \dfrac{1}{n} \sum_{k=1}^{n} y_k^* - \overline{y}$ et

$\dfrac{1}{n} \sum_{k=1}^{n} y_k^* = \dfrac{1}{n} \sum_{k=1}^{n} (\hat{a}\, x_k + \hat{b}) = \hat{a} \dfrac{1}{n} \sum_{k=1}^{n} x_k + \hat{b} = \hat{a}\, \overline{x} + \hat{b} = \overline{y}$ car la droite de régréssion passe par le point $(\overline{x},\, \overline{y})$.

3.2.4 Equation de la variance

- On appelle somme des carrés totale la quantité positive : $S_T = \sum_{k=1}^{n} (y_k - \overline{y})^2 = n\, V(y)$
- On appelle somme des carrés de la régression la quantité positive : $S_R = \sum_{k=1}^{n} (y_k^* - \overline{y})^2$.
- On appelle somme des carrés résiduellee la quantité positive : $S_E = \sum_{k=1}^{n} (y_k - y_k^*)^2$.
- On appelle équation de la variance : $S_T = S_R + S_E$. En effet

$S_T = \sum_{k=1}^{n} (y_k - \overline{y})^2 = \sum_{k=1}^{n} (y_k - y_k^* + y_k^* - \overline{y})^2 = \sum_{k=1}^{n} (y_k - y_k^*)^2 + \sum_{k=1}^{n} (y_k^* - \overline{y})^2 + 2 \sum_{k=1}^{n} (y_k - y_k^*)(y_k^* - \overline{y})$.

Montrons que $\sum_{k=1}^{n} (y_k - y_k^*)(y_k^* - \overline{y}) = 0$. En remplaçant y_k^* par $\hat{a}\, x_k + \hat{b}$, on a

$\sum_{k=1}^{n} (y_k - y_k^*)(y_k^* - \overline{y}) = \sum_{k=1}^{n} (y_k - \hat{a}\, x_k - \hat{b})(\hat{a}\, x_k + \hat{b} - \overline{y})$, en remplaçant $\hat{b}$ par $\overline{y} - \hat{a}\overline{x}$, on obtient

$\sum_{k=1}^{n} ((y_k - \overline{y}) - \hat{a}(x_k - \overline{x}))\hat{a}(x_k - \overline{x}) = \hat{a} \sum_{k=1}^{n} (y_k - \overline{y})(x_k - \overline{x}) - \hat{a}^2 \sum_{k=1}^{n} (x_k - \overline{x})^2 = \hat{a} Cov(x,\, y) - \hat{a}^2 V(x)$

et en remplaçant $\hat{a}$ par $\dfrac{Cov(x,\, y)^2}{V(x)}$ on trouve $\dfrac{n\, Cov(x,\, y)^2}{V(x)} - \dfrac{Cov(x,\, y)^2}{V(x)^2}\, n\, V(x) = 0$.

- On appelle coefficient de détermination la quantité positive : $R^2 = \dfrac{S_R}{S_T} = \dfrac{\sum_{k=1}^{n} (y_k^* - \overline{y})^2}{\sum_{k=1}^{n} (y_k - \overline{y})^2}$.

On a

$$0 \leq R^2 \leq 1$$

En effet $0 \leq S_R \leq S_R + S_E = S_T$ En divisant le tout par S_T On a le résultat.

Le coefficient de détermination R^2 nous donne le pourcentage expliqué par la régression.

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 on peut approché Y par la droite $a\,X + b$ avec

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

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

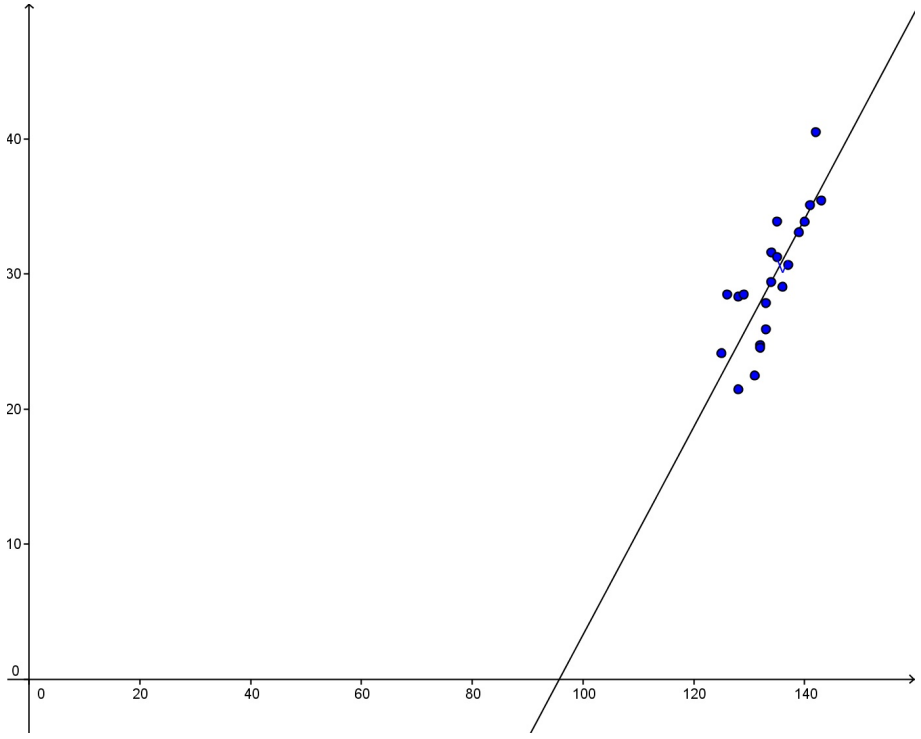

FIGURE 3.3 – nuage de points et droite de régression