FACULTÉ DES SCIENCES ET TECHNIQUES-ERRACHIDIA

Polycopié

Statistiques descriptives

Filière M.I.P., S3

Module M136

Pr. M. R. Sidi Ammi

Département de Mathématiques

2015/2016

Table des matières

Table des matières **1**

1 Calcul Statistique **2**

1.1 Objet et langage de la Statistique . . . 2

1.1.1 Introduction . . . 2

1.1.2 Population - Caractère . . . 2

1.1.3 Effectif, série (ou distribution) statistique . . . 3

1.2 Distributions statistiques à un caractère . . . 4

1.2.1 Définitions et exemples . . . 4

1.2.2 Présentation des résultats : Représentations graphiques . . . 5

1.2.3 Paramètres de position . . . 6

1.2.4 Paramètres de dispersion . . . 11

1.3 Distributions statistiques à deux caractères . . . 13

1.3.1 Définitions et exemples . . . 13

1.3.2 Distributions marginales . . . 15

1.3.3 Fréquences conditionnelles, distributions conditionnelles, indépendance . . . 16

1.3.4 Coefficient de corrélation . . . 17

2 Ajustement linéaire **20**

2.1 Introduction . . . 20

2.2 Alignement statistique . . . 20

2.2.1 Méthode des moindres carrés . . . 22

2.2.2 Généralisation . . . 25

2.2.3 Exercices . . . 27

CHAPITRE 1

Calcul Statistique

1.1 Objet et langage de la Statistique

1.1.1 Introduction

La statistique est un ensemble de méthodes scientifiques basées sur la collecte, l'organisation, la présentation de données ainsi que sur la modélisation et la construction de résumés numériques et qui permettent de décrire et d'analyser des phénomènes repérés par des éléments de même nature, susceptibles d'être dénombrés et/ou classés.
Le rôle d'explication et de prévision appartient à l'utilisateur et non à la statistique, qui n'est qu'un outil d'investigation.
Le calcul statistique a double objectif :

1^{er} but : A partir de données brutes en grand nombre, on peut dégager avec un minimum d'effort, un certain nombre de renseignements qualitatifs ou quantitatifs permettant de visualiser cette statistique avec une bonne précision. Cela afin de pouvoir la comparer à d'autres statistiques du même type (Statistique Descriptive).

$2^{\text{ème}}$ but : La statistique ayant été effectuée sur un échantillon, nous permet d'extrapoler les résultats partiels en vue de déduire des précisions globales (Statistique Inférentielle).
On prendra garde à ne pas confondre **la statistique** - ensemble de méthodes scientifiques - et **les statistiques**, terme désignant les résultats numériques d'une enquête ou d'une série de mesures (après traitement éventuel par la statistique).
Ainsi, nous allons nous limiter dans ce cours au 1^{er} but. Nous le traitons en s'appuyant aussi souvent que possible sur des exemples.

1.1.2 Population - Caractère

Toute étude statistique considère au départ un ensemble Ω, appelé **population** (personnes, villes, voitures, virus,...). Tout sous-ensemble de la population Ω est appelé **échantillon**. Cette terminologie est issue de la démographie, première science à avoir développé des méthodes statistiques ; il est cependant clair que la population que l'on envisage en statistique dépend du domaine que l'on traite, et peut donc aussi bien être constituée d'êtres humains que d'animaux, d'objets, voire d'événements.
Chaque élément de la population P est appelé **individu** et consiste à observer et étudier un même aspect sur chaque individu nommé **caractère** (nombre d'enfants, consommation, marques de voitures,...).
Plus généralement, on appelle caractère toute application X de la population Ω dans un ensemble E, dont les éléments x sont appelés modalités du caractère X (ou valeurs du caractère).

$$\begin{array}{rl} X \ : & \Omega \longrightarrow E \\ & w \longmapsto X(w) := x, \end{array}$$

X : le caractère
Ω : la population
w : individu
E : l'ensemble des modalités
$X(w) := x$: modalité.

Il existe deux types de caractère :

1. **Caractère quantitatif** : mesurable c'est-à-dire auquel on peut associer un nombre (le nombre d'enfants, la taille, la masse, la longueur, le volume). Un caractère quantitatif est souvent appelé également **variable statistique**.
   On distingue alors deux types de caractère quantitatif :
   - **Caractère quantitatif discret** : qui ne peut prendre qu'un nombre fini de valeurs (de modalités) isolées ($X(\Omega) = \{0, 1, 2, 3, 4\}$). Par exemple : nombre de voitures par famille.
   - **Caractère quantitatif continu** : qui, théoriquement, peut prendre toutes les valeurs d'un sous-ensemble non dénombrable de $\mathbb{R}$, (en pratique un intervalle de $\mathbb{R}$). Ses valeurs sont alors regroupées en classes ($X(\Omega) = [0, 5[\cup[5, 10[\cup[10, 15[\cup[15, 20[$). Par exemple : la taille d'un individu, le temps passé devant la télé.
2. **Caractère qualitatif** : Ses modalités sont des qualités ($X(\Omega) = \{\text{blanc}, \text{bleu}, \text{jaune}, \text{rouge}, \text{beige}\}$). Par exemple : sexe, profession, nationalité, marque de voiture.

Exemple 1.1 *Etude du nombre d'enfants des couples d'un quartier donné.*

Population : Ensembles des couples du quartier,
Individu : couple,
Caractère : Nombre d'enfants,
Type : Quantitatif discret.

Exemple 1.2 *Etude des marques de voitures d'une ville donnée.*

Population : Ensembles des voitures de la ville,
Individu : voiture,
Caractère : Marque,
Type : Qualitatif.

Remarque 1.1 *Toute statistique qualitative peut se transformer en une statistique quantitative à l'aide d'un codage des valeurs possibles du caractère. Par exemple : 1 :masculin et 2 :féminin est le codage usuel du sexe, le code postal : codage de lieux géographiques.*

1.1.3 Effectif, série (ou distribution) statistique

Soient X est un caractère quantitatif discret et Ω une population finie. Si on pose $X(\Omega) = \{x_1, x_2, ..., x_i, ..., x_p\}$ alors pour chaque valeur x_i de modalité (du caractère) constatée, on détermine le nombre d'individus n_i ayant présenté cette valeur du caractère, nombre appelé effectif associé à la modalité. L'ensemble des couples (x_i, ni) ((modalité, effectif)) ainsi déterminé est parfois appelé **distribution statistique** ou **série statistique** ou encore **variable statistique**. On dit alors que l'on a effectué un **regroupement** des données brutes.
Une **série statistique à un caractère** ou simple, ou à une dimension, est obtenue lorsque nous nous intéressons à un caractère élémentaire, dont l'ensemble des modalités $X(\Omega)$ est un sous-ensemble de $\mathbb{R}$ s'il est quantitatif.
Une **série statistique à deux caractères** ou double est obtenue lorsque à chaque individu sont associés deux caractères élémentaires, plus précisément un couple de caractères élémentaires, ou encore un caractère à valeurs dans le produit cartésien $\mathbb{R}^2$ c'est-à-dire $X(\Omega) \subset \mathbb{R}^2$.
La différence entre l'étude de deux caractères simples sur la même population et d'un caractère double sur cette même population peut paraître artificielle : elle est cependant essentielle. En effet, dans le cas d'une série double, nous nous intéressons pour chaque individu au couple (x,y) de réponses et nous effectuons le regroupement des données par rapport à ces couples, alors que dans le cas de l'étude des deux séries simples associées nous effectuons le regroupement des données séparément sur chacun des deux caractères X et Y ; nous obtenons alors des résultats plus concis, mais au prix d'une perte d'information.
Prenons un exemple concret :

Exemple 1.3 *Nous effectuons un sondage auprès de nos étudiants en leur demandant leur note de mathématique au baccalauréat et le nombre de redoublements au cours de leur scolarité primaire et secondaire. Les résultats bruts obtenus sont les*

suivants :

14-0	*12-1*	*11-0*	*10-2*	*15-0*	*13-1*	*11-2*	*10-3*
11-0	*12-1*	*13-1*	*14-0*	*13-0*	*11-1*	*12-0*	*13-1*

Soit X le caractère "note au bac" et Y le caractère "nombre de redoublements". Les tableaux statistiques regroupant les données de X et de Y sont :

x_i	*10*	*11*	*12*	*13*	*14*	*15*
n_i	*2*	*4*	*3*	*4*	*2*	*1*

et

y_i	*0*	*1*	*2*	*3*
n_i	*7*	*6*	*2*	*1*

alors que le tableau statistique regroupant les données du couple (X ; Y) est :

X$\diagdown$*Y*	*0*	*1*	*2*	*3*
10	*0*	*0*	*1*	*1*
11	*2*	*1*	*1*	*0*
12	*1*	*2*	*0*	*0*
13	*1*	*3*	*0*	*0*
14	*2*	*0*	*0*	*0*
15	*1*	*0*	*0*	*0*

Nous pouvons remarquer que si les tableaux statistiques des variables X et Y ne permettent pas de reconstruire le tableau du couple (X,Y), ce dernier par contre permet de retrouver les tableaux de X et de Y en effectuant la somme des effectifs par colonne et par ligne :

X$\diagdown$*Y*	*0*	*1*	*2*	*3*	
10	*0*	*0*	*1*	*1*	*2*
11	*2*	*1*	*1*	*0*	*4*
12	*1*	*2*	*0*	*0*	*3*
13	*1*	*3*	*0*	*0*	*4*
14	*2*	*0*	*0*	*0*	*2*
15	*1*	*0*	*0*	*0*	*1*
	7	*6*	*2*	*1*	

Lorsque l'on étudie la loi du couple (X,Y) les distributions de X et de Y sont appelées **distributions marginales**, la distribution du couple étant la **distribution conjointe**.
L'intérêt de la distribution conjointe par rapport aux distributions marginales est de permettre l'étude de la **corrélation** entre les deux caractères X et Y, c'est-à-dire de savoir s'il existe un rapport (et éventuellement quel rapport) entre la note au bac et le nombre de redoublements antérieurs.

1.2 Distributions statistiques à un caractère

1.2.1 Définitions et exemples

Soit $X : \Omega \longrightarrow \mathbb{R}$ une statistique à un caractère.

(i) **Cas où X est un caractère quantitatif discret :** Posons $X(\Omega) = \{x_1, x_2, ..., x_i, ..., x_p\}$ avec la convention $x_1 < x_2 < ... < x_i < ... < x_p$. Soit $i \in \{1, 2, ..., p\}$.

- On appelle **effectif** de la valeur x_i, le cardinal n_i de l'ensemble $X^{-1}(\{x_i\})$.
- On appelle **effectif cumulé** en x_i, le nombre $N_i = \sum_{j=1}^{i} n_j$.
- On appelle **effectif total** de la série statistique (x_i, n_i), le nombre $N = \sum_{i=1}^{p} n_i$.
- On appelle **fréquence** de la valeur x_i, le nombre $f_i = \frac{n_i}{N}$.
- On appelle **fréquence cumulée** en x_i, le nombre $F_i = \sum_{j=1}^{i} f_j$.

– On appelle **pourcentage** de la valeur x_i, le nombre $p_i = 100 f_i$.

(ii) **Cas où X est un caractère quantitatif continu :** Notons $I_i = [x_{i-1}, x_i[$ et posons $X(\Omega) = \bigcup_{i=1}^{p} I_i$ avec $I_i \cap I_j = \emptyset$ pour tout $i\,, j \in \{1, 2, ..., p\}$ avec $i \neq j$.

– Chaque intervalle $I_i = [x_{i-1}, x_i[,\ i \in \{1, 2, ..., p\}$, est appelé **classe**.

– On appelle **effectif de la classe** $I_i = [x_{i-1}, x_i[$, le cardinal de $X^{-1}(I_i)$.

– On appelle **effectif cumulé** en $I_i = [x_{i-1}, x_i[$, le cardinal de $X^{-1}(\bigcup_{j=1}^{i} I_j)$.

– La série statistique (I_i, n_i) est souvent notée (c_i, n_i) où $c_i = \frac{x_{i-1}+x_i}{2}$ désigne le **centre de la classe** $I_i = [x_{i-1}, x_i[$.

Exemple 1.4 *On donne la répartition des notes obtenues à un partiel :*

Note x_i	*3*	*7*	*8*	*10*	*12*	*15*	*16*
Effectif n_i	*2*	*6*	*7*	*8*	*4*	*2*	*1*

Population : Ensembles des étudiants,
Individu : étudiant,
Caractère : la note du partiel,
Type : Quantitatif discret.
En faisant les calculs, on trouve

Note	x_i	*3*	*7*	*8*	*10*	*12*	*15*	*16*
Effectif	n_i	*2*	*6*	*7*	*8*	*4*	*2*	*1*
Effectif cumulé	N_i	*2*	*8*	*15*	*23*	*27*	*29*	*N=30*
Fréquence	f_i	$\frac{2}{30}$	$\frac{6}{30}$	$\frac{7}{30}$	$\frac{8}{30}$	$\frac{4}{30}$	$\frac{2}{30}$	$\frac{1}{30}$
Fréquence cumulée	F_i	$\frac{2}{30}$	$\frac{8}{30}$	$\frac{15}{30}$	$\frac{23}{30}$	$\frac{27}{30}$	$\frac{29}{30}$	*1*
Pourcentage	p_i	*6.66%*	*20%*	*23.33%*	*26.66%*	*13.33%*	*6.66%*	*3.33%*

Ainsi, le pourcentage des étudiants qui ont réussi est : $p_4 + p_5 + p_6 + p_7 = 100\frac{8+4+2+1}{30} = 50\%$.

Exemple 1.5 *La répartition des salaires d'une entreprise est la suivante :*

Salaire en DH	*Classe* I_i	*[400,600[*	*[600,800[*	*[800,1000[*	*[1000,1200[*
Centre de I_i	c_i	*500*	*700*	*900*	*1100*
Effectif	n_i	*20*	*96*	*52*	*17*
Effectif cumulé	N_i	*20*	*116*	*168*	*N=185*
Fréquence	f_i	$\frac{20}{185}$	$\frac{96}{185}$	$\frac{52}{185}$	$\frac{17}{185}$
Fréquence cumulée	F_i	$\frac{20}{185}$	$\frac{116}{185}$	$\frac{168}{185}$	*1*
Pourcentage	p_i	*10.81%*	*51.89%*	*28.11%*	*9.19%*

1.2.2 Présentation des résultats : Représentations graphiques

Lorsque l'on a observé une série statistique, il est souvent souhaitable de présenter les résultats sous forme graphique. Un tableau statistique, aussi clair soit-il, n'est jamais aussi parlant qu'une représentation graphique, même si par ailleurs il recèle davantage d'informations. Dans le cas des statistiques, on parlera souvent de diagrammes au lieu de représentation graphique.

Diagramme en batons, histogramme et diagramme en secteurs. Reprenons l'exemple 1.4, le diagramme en batons ci-dessous (figure 1.1) présente l'effectif en fonction des modalités. L'histogramme ci-dessous (figure 1.2), met en relief l'effectif en fonction des modalités du caractère dans l'exemple 1.5. Dans un diagramme en secteurs (figure 1.3) (appelé encore camembert), les effectifs des différentes classes sont représentés par des secteurs d'angles proportionnels aux effectifs. Si on reprend l'exemple 1.5, on peut avoir une représentation sous forme de diagramme en secteurs :

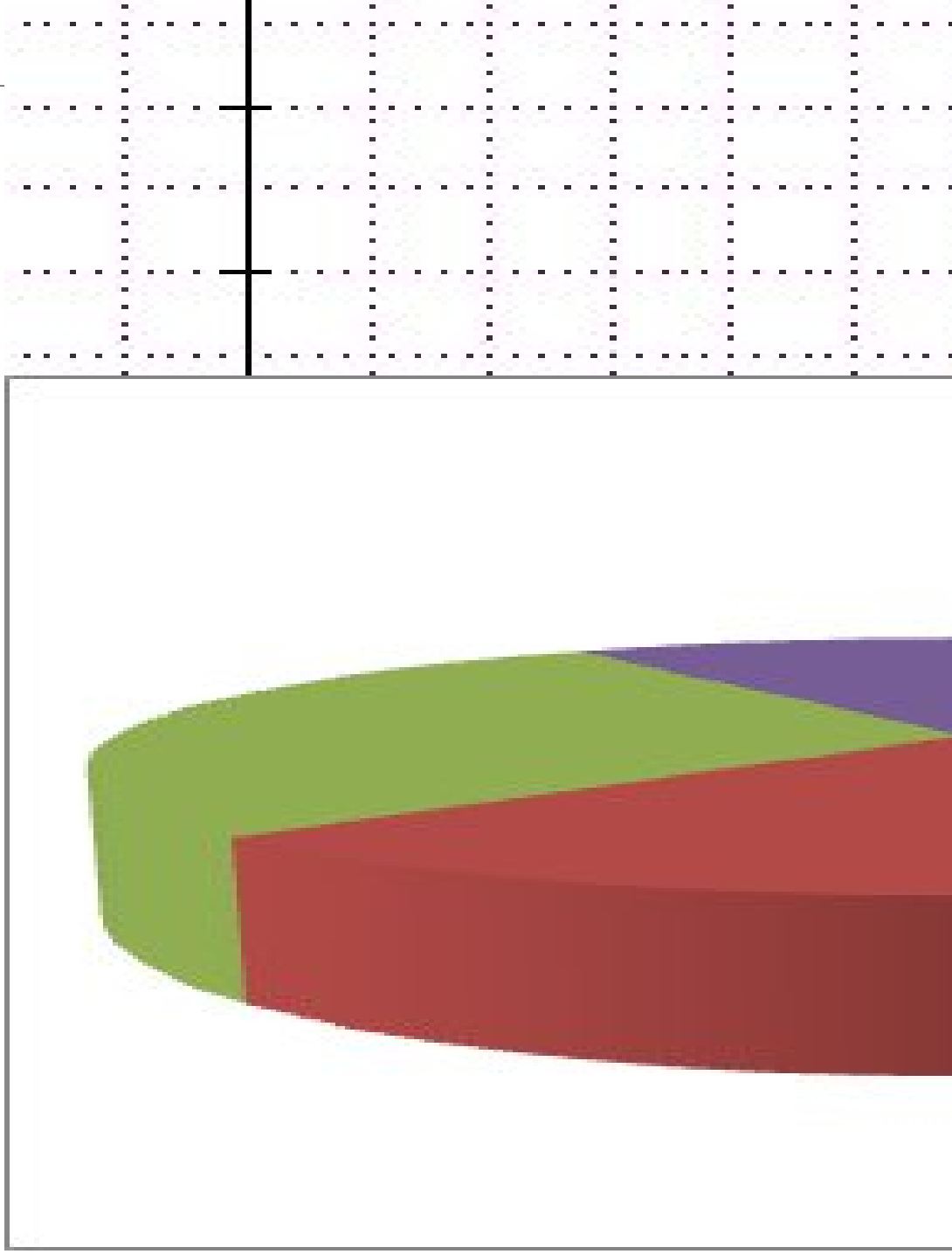


FIGURE 1.3 – Diagramme en secteurs.

1.2.3 Paramètres de position

Appelés également caractéristiques de position ou de tendance centrale, les paramètres de position tentent de donner une information sur la valeur de la modalité "autour de laquelle se situent les autres modalités" (d'où le terme de tendance "centrale").
Cette notion suppose donc que les modalités constituent un ensemble ordonné, ou pour le moins que la notion de "distance" entre deux modalités ait un sens.
Il est bien évident que seuls les caractères quantitatifs permettent de donner un sens précis à ces notions intuitives.

Cas d'un caractère quantitatif discret

Une série statistique discrète $(x_i, n_i)_{i\in\{1,2,\ldots,p\}}$ est dite ordonnée si $i < j \Rightarrow x_i < x_j$.

Définition 1.1 *Soit $X = (x_i, n_i)_{i\in\{1\dots p\}}$ est une série statistique ordonnée discrète.*

1. *On appelle mode, toute valeur de la modalité dont l'effectif est maximum.*
   Autrement dit, le mode de la série statistique est toute modalité x_i telle que :

$$n_i = \max_{j\in\{1\dots p\}} n_j.$$

2. *On appelle moyenne arithmétique de la série statistique, le nombre réel noté $\overline{X}$, ou $E(X)$, définit par :*

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i, \qquad \text{où } N = \sum_{i=1}^{p} n_i.$$

3. *On appelle médiane de la série statistique, toute modalité x_i du caractère telle que :*

$$\sum_{j/x_j<x_i} n_j \leq \frac{N}{2} \quad \text{et} \quad \sum_{j/x_j>x_i} n_j \leq \frac{N}{2}.$$

Exemple 1.6 *Reprenons l'exemple 1.4,*

1. *le mode est la modalité $x_4 = 10$.*
2. *La moyenne arithmétique est $\overline{X} \backsimeq 9.26$.*
3. *La médiane est la modalité x_3, en effet on a $2 + 6 = 8 \leq 15$ et $8 + 4 + 2 + 1 = 15$.*

Exemple 1.7 *On considère la série statistique suivante :*

x_i	3	5	6	7	8	9
n_i	1	2	1	1	2	3

L'effectif total est $N = 10$.

- *Le mode est : $x_6 = 9$.*
- *La moyenne arithmétique est $\overline{X} = 6.9$.*
- *La modalité $x_4 = 7$ est une médiane, car $1 + 2 + 1 = 4 \leq \frac{N}{2} = 5$, et $2 + 3 = \frac{N}{2}$.*
- *$x_5 = 8$ est aussi une médiane.*

Remarque 1.2 *Une série statistique peut avoir plus qu'une valeur modale et plus qu'une médiane.*

On peut tout de même caractériser la médiane par son effectif cumulé.

Proposition 1.1 *Soit $(x_i, n_i)_{i\in\{1,2,\dots,p\}}$ une série statistique discrète ordonnée. La plus petite modalité dont l'effectif cumulé est supérieur ou égale à la moitié de l'effectif total est une médiane.*

Exemple 1.8 *Soit la distribution statistique suivante :*

x_i	2	4	5	7	10	12	16
n_i	4	3	4	6	8	3	2
N_i	4	7	11	17	25	28	N=30

- *Le mode est : $x_5 = 10$.*
- *La moyenne arithmétique est : $\overline{X} \backsimeq 7.66$.*
- *Il est clair que la plus petite modalité dont l'effectif cumulé dépasse la moitié de l'effectif total est $x_4 = 7$. Soit $N_4 = 17 \geq \frac{N}{2} = 15$. Ainsi, $x_4 = 7$ est une médiane.*

Cas d'un caractère quantitatif continu (distribution groupée) :

Définition 1.2 *Soit $X = (I_i = [a_{i-1}, a_i[, n_i)_{i\in\{1,\dots,p\}}$ une série statistique regroupée en classes.*

1. *On appelle classe modale de la série statistique, toute classe dont le rapport $\dfrac{\text{effectif}}{\text{largeur de la classe}}$ est maximum.*

2. *On appelle moyenne arithmétique de la série statistique, le nombre réel noté* $\overline{X}$, $E(X)$, *définit par*

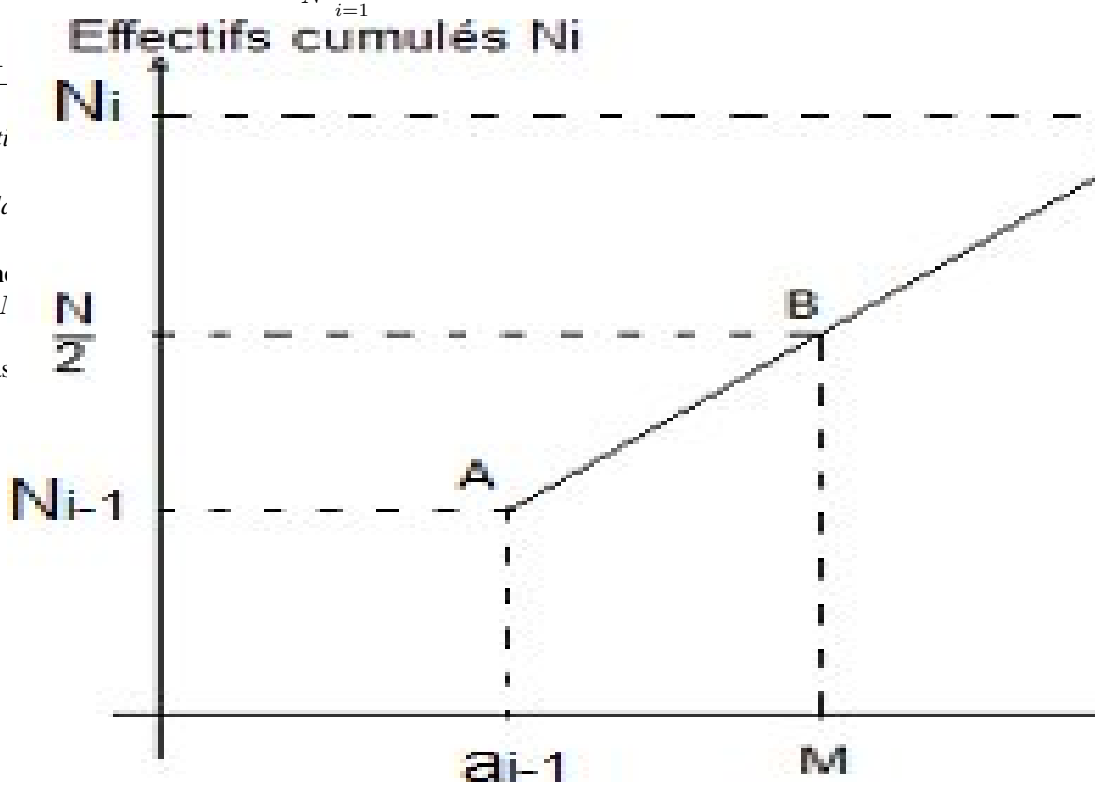
$$E(X) = \frac{1}{N}\sum_{i=1}^{p} n_i c_i.$$

où $N = \sum_{i=1}^{p} n_i$ *et* $c_i = \frac{a_{i-1} +$

3. *La médiane d'une série statisti*

Remarque 1.3 *Si les largeurs des cl*

Détermination de la valeur médian
Graphiquement, la valeur médiane l statistique.
Notons N_i l'effectif cumulé de la clas



Les points A, B et C étant alignés, ainsi on a $det(\overrightarrow{AB}, \overrightarrow{AC}) = 0$, et nous obtenons alors

$$\frac{M - a_{i-1}}{a_i - a_{i-1}} = \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}.$$

Exemple 1.9 *Reprenons l'exemple 1.5,*

Salaire en DH *Classe* I_i		*[400,600[*	*[600,800[*	*[800,1000[*	*[1000,1200[*
Centre de I_i	c_i	*500*	*700*	*900*	*1100*
Effectif	n_i	*20*	*96*	*52*	*17*
Effectif cumulé	N_i	*20*	*116*	*168*	*N=185*
Fréquence	f_i	$\frac{20}{185}$	$\frac{96}{185}$	$\frac{52}{185}$	$\frac{17}{185}$
Fréquence cumulée	F_i	$\frac{20}{185}$	$\frac{116}{185}$	$\frac{168}{185}$	*1*
Pourcentage	p_i	*10.81%*	*51.89%*	*28.11%*	*9.19%*

– *la classe modale est* $I_2 = [600, 800[$.
– *La moyenne arithmétique est :* $E(X) \backsimeq 771.35$.
– *La médiane* M *:* $N_1 = 20 \leq \frac{N}{2} = 92.5 \leq N_2 = 116 \Rightarrow M \in [600, 800[$. *Ainsi on a*

$$\frac{M - 600}{800 - 600} = \frac{92.5 - 20}{116 - 20}.$$

Soit $M \backsimeq 751.042$.

Exemple 1.10 *Un jury a attribué des notes regroupées dans le tableau suivant :*

Notes x_i	*[0,5[*	*[5,8[*	*[8,12[*	*[12,15[*	*[15,20[*
Effectifs n_i	*10*	*8*	*12*	*11*	*9*

Pour chercher les paramètres de position, on doit completer le tableau pour avoir :

Notes	x_i	*[0,5[*	*[5,8[*	*[8,12[*	*[12,15[*	*[15,20[*
Effectifs	n_i	*10*	*8*	*12*	*11*	*9*
Effectifs cumulés	N_i	*10*	*18*	*30*	*41*	*N=50*
Centres	c_i	*2.5*	*6.5*	*10*	*13.5*	*17.5*

- *la classe modale est* $I_3 = [8, 12[$.
- *La moyenne arithmétique est :* $E(X) = 10.06$.
- *La médiane* M : *On a* $N_2 = 18 \leq \frac{N}{2} = 25 \leq N_3 = 30 \Rightarrow M \in [8, 12[$. *Alors,*

$$\frac{M-8}{12-8} = \frac{25-18}{30-18}.$$

Soit $M \backsimeq 10.33$.

Quantiles d'ordre α **(**$0 < \alpha < 1$**) :**

Définition 1.3 *On appelle quantile d'ordre* α *d'une série statistique* $(x_i, n_i)_{i=1,n}$, *la modalité de cette série pour laquelle l'effectif cumulé représente le quotient* α *de l'effectif total.*

Ainsi, les quantiles sont donc des nombres qui partagent la série statistique en des parties qui ont toutes "sensiblement" le même nombre de termes.
On remarque immédiatement que la médiane est un quartile particulier, elle correspond au quantile d'ordre $\frac{1}{2}$ puisque la médiane désigne la valeur du caractère ayant un effectif cumulé de 50%.
Cas particuliers : quartiles, deciles, centiles.

- Les quartiles sont des valeurs du caractères qui correspondent aux valeurs $\alpha = \frac{1}{4}$, $\alpha = \frac{1}{2}$ et $\alpha = \frac{3}{4}$.
  1. Le premier quartile noté Q_1 est la valeur du caractère pour un effectif cumulé de 25% de l'effectif total.
  2. Le deuxième quartile noté Q_2 est la valeur du caractère pour un effectif cumulé de 50% de l'effectif total, c'est en fait la médiane.
  3. Le troisième quartile noté Q_3 est la valeur du caractère pour un effectif cumulé de 75% de l'effectif total.
- Si $\alpha = \frac{i}{10}N$ avec $i \in \{1, 2, ..., 9\}$ alors on parle du $i^{\text{ème}}$ decile.
- Si $\alpha = \frac{i}{100}N$ avec $i \in \{1, 2, ..., 99\}$ alors on parle du $i^{\text{ème}}$ centile.

Définition 1.4 *L'intervalle interquartile est l'intervalle* $[Q_1, Q_3]$
L'écart interquartile est le nombre $Q_3 - Q_1$. *C'est la largeur de l'intervalle interquartile.*

Remarque 1.4 *L'intervalle interquartile élimine les valeurs extrêmes, c'est peut être un avantage, mais en revanche il ne compte que* 50% *de l'effectif total de la série statistique.*

Exemple 1.11 *Cas d'une distribution a caractère discret :*
On considère la distribution statistique suivante :

x_i	*7*	*8*	*9*	*10*	*11*	*12*	*13*
n_i	*2*	*3*	*7*	*10*	*6*	*2*	*1*
N_i	*2*	*5*	*12*	*22*	*28*	*30*	$N = 31$

- $\frac{1}{4}N = 7.75$. *Ainsi, le premier quartile est* $Q_1 = 9$.
- $\frac{1}{2}N = 15.5$. *Ainsi, le deuxième quartile (la médiane) est* $Q_2 = 10$.
- $\frac{3}{4}N = 23.25$. *Ainsi, le troisième quartile est* $Q_3 = 11$.
- *L'écart interquartile est* $Q_3 - Q_1 = 11 - 9 = 2$.
- $\frac{1}{10}N = 3.1$. *Ainsi, le premier decile est* $D_1 = 8$.
- $\frac{4}{10}N = 12.4$. *Ainsi, le quatrième decile est* $D_4 = 10$.
- $\frac{1}{100}N = 0.31$. *Ainsi, le premier centile est* $C_1 = 7$.
- $\frac{9}{100}N = 2.79$. *Ainsi, le neuvième centile est* $C_9 = 8$.

FIGUI

– $\frac{25}{100}N = 7.75$. *Ainsi, le vingt-ci*

Exemple 1.12 *Cas d'une distribution*
Un test de mémorisation a été effectue

Age
Nomb
Effe

– $\frac{1}{4}N = 12.5$. *Ainsi, le premier q*
– $\frac{1}{2}N = 25$. *Ainsi, le deuxième q*
– $\frac{3}{4}N = 37.5$. *Ainsi, le troisième*
– **Calcul des quartiles**. *Comme p*

$$\left\{$$

Avec
– N_i *l'effectif cumulé correspo*
– N_a *l'effectif cumulé avant la*
– N_b *l'effectif cumulé jusqu'à l*

– $Q_1 = 41 + \dfrac{50-41}{25-12}(12.5 - 12$
– $Q_2 = 41 + \dfrac{50-41}{25-12}(25 - 12)$
– $Q_3 = 51 + \dfrac{60-51}{50-25}(37.5 - 25$
– *L'écart interquartile est* $Q_3 - ($

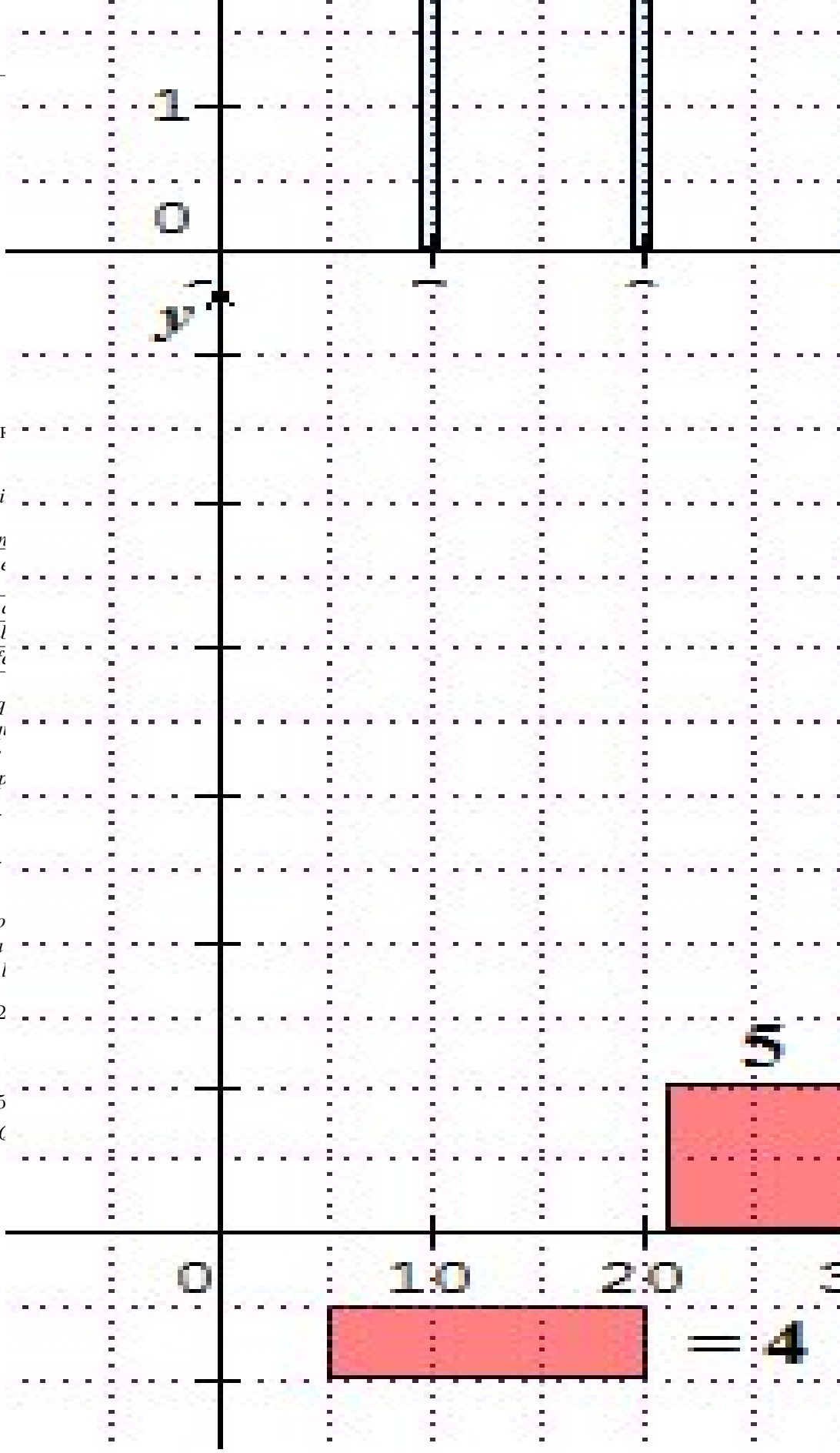


FIGURE 1.5 – Histogramme avec médiane et quartiles.

Proposition 1.2 *Soient $a, \lambda \in \mathbb{R}$ et $X = (x_i, n_i)$ une série statistique. Nous notons $X + a$ la série $(x_i + a, n_i)$ et λX la série $(\lambda x_i, n_i)$. Alors*

$$E(X + a) = E(X) + a$$

et

$$E(\lambda X) = \lambda E(X).$$

Remarque 1.5 *Les propriétés précédentes expriment, dans le cas de mesures de grandeurs physiques, que la moyenne arithmétique ne dépend pas ni choix de l'origine ni de l'unité. Ainsi, une moyenne de températures exprimées en degrés Celsius (Kelvin) sera obtenue en degrés Celsius (Kelvin). Ces propriétés sont également utilisées pour faciliter les calculs et l'entrée des données dans un programme de calcul. Si nous considérons par exemple la série statistique X donnant la taille d'un échantillon d'étudiants :*

X	*163*	*170*	*171*	*175*	*178*	*180*	*181*	*182*	*185*
n_i	*1*	*1*	*3*	*1*	*3*	*2*	*1*	*2*	*6*

on peut considérer la variable $X - 180$ par exemple. Nous obtenons alors le tableau suivant :

X	*163*	*170*	*171*	*175*	*178*	*180*	*181*	*182*	*185*
$X - 180$	*-17*	*-10*	*-9*	*-5*	*-2*	*0*	*1*	*2*	*5*
n_i	*1*	*1*	*3*	*1*	*3*	*2*	*1*	*2*	*6*

D'où

$$E(X - 180) = \frac{1}{20}(-17 - 10 - 27 - 5 - 6 + 0 + 1 + 4 + 30) = -\frac{30}{20} = -1.5.$$

Ainsi,

$$E(X) = 180 - 1.5 = 178.5$$

1.2.4 Paramètres de dispersion

Dans le paragraphe précédent, nous avons quantifié l'aspect "moyen" d'une série statistique. Il est bien évident que des séries ayant la même moyenne et la même médiane peuvent être très différemment étalées. Les paramètres de dispersion que nous allons maintenant définir permettent d'apprécier l'importance de la dispersion d'une série statistique autour de sa moyenne ou de sa médiane.

Étendue, Écart-moyen, Variance, Écart-type

Définition 1.5 *On appelle étendue d'une distribution statistique quantitative $X = (x_i, n_i)_{i\in\{1,2,\dots,p\}}$ la différence entre la plus grande et la plus petite valeur observée, i.e.* $\max x_i - \min x_i$.

Bien que très "primitif", cet indice de dispersion n'est pas à négliger.

Définition 1.6 *Soit $X = (x_i, n_i)_{i\in\{1,2,\dots,p\}}$ une série statistique de moyenne arithmétique $\overline{X}$ et d'effectif total $N = \sum_{i=1}^{p} n_i$.*

1. *On appelle écart moyen de la série statistique, le nombre réel positif noté $e(X)$ définit par :*

$$e = \frac{1}{N}\sum_{i=1}^{p} n_i \times |x_i - \overline{X}|.$$

2. *On appelle variance de la série statistique, le nombre réel positif noté $V(X)$ définit par :*

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

3. *On appelle écart-type de la série statistique, le nombre réel positif noté $\sigma(X)$ définit par :*

$$\sigma(X) = \sqrt{V(X)}.$$

Si $X = \big(I_i = [a_{i-1}, a_i[, n_i\big)_{i\in\{1,\dots,p\}}$, alors on remplace I_i par $c_i = \dfrac{a_{i-1} + ai}{2}$ dans les définitions ci-dessus.

Remarque 1.6 – *L'écart moyen et la variance calculent la moyenne des écarts et des carrés des écarts respectivement par rapport à la moyenne arithmétique.*
– *L'écart moyen et la variance mesurent la dispersion des valeurs du caractère de la moyenne arithmétique.*
– *l'introduction de la variance est justifier par le fait qu'elle réalise le minimum de la fonction*

$$g(t) = \frac{1}{N}\sum_{i=1}^{p} n_i \times (x_i - t)^2.$$

En fait

$$V(X) = \min_t g(t).$$

– *La plus part des valeurs du caractère se trouvent dans l'intervalle* $[\overline{X} - \sigma(X), \overline{X} + \sigma(X)]$. *Ainsi, d'autant que* $\sigma(X)$ *est petit d'autant que les valeurs du caractère s'approchent de la moyenne arithmétique de la série statistique. Par conséquent, l'écart-type* $\sigma(X)$ *est le plus signifiant des paramètres des dispersion.*

Exemple 1.13 *La distribution des notes de mathématiques de deux étudiants* A *et* B *durant une année est donnée dans le tableau suivant :*

Notes x_i		*12*	*14*	*15*	*16*	*17*	*18*
effectif	*A*	*3*	*1*	*1*	*1*	*3*	*1*
n_i	*B*	*0*	*2*	*6*	*2*	*0*	*0*

On calcul
– *les moyennes arithmétiques des deux étudiants :* $\overline{X}_A = \overline{X}_B = 15$.
– *L'écart moyen de l'étudiant* A *est :* $e_A = 2$ *et celui de* B *est* $e_B = 0.4$.
On remarque que $e_B < e_A$, *cela s'explique par le fait que le travail de l'étudiant* B *est plus régulier que celui de l'étudiant* A.

Proposition 1.3 *Soient* $a, b \in \mathbb{R}$ *et* X *une série statistique. On a*
– $V(X) = \overline{X^2} - \overline{X}^2$ *(Relation de Kœnig-Huyghens).*
– $V(X+b) = V(X)$.
– $V(aX) = a^2 V(X)$.

Preuve.
–

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

–

$$\begin{aligned} V(X+b) &= \overline{(X+b)^2} - \overline{X+b}^2 \\ &= \overline{X^2 + 2bX + b^2} - (\overline{X} - b)^2 \\ &= \overline{X^2} + 2b\overline{X} + b^2 - \overline{X}^2 - 2b\overline{X} - b^2 \\ &= \overline{X^2} - \overline{X}^2 \\ &= V(X). \end{aligned}$$

–

$$\begin{aligned} V(aX) &= \overline{(aX)^2} - \overline{aX}^2 \\ &= \overline{a^2X^2} - a^2\overline{X}^2 \\ &= a^2\overline{X^2} - a^2\overline{X}^2 \\ &= a^2 V(X). \end{aligned}$$

Exemple 1.14 *Reprenons l'exemple 1.13. Pour l'étudiant A on a le tableau suivant :*

Notes x_i	*12*	*14*	*15*	*16*	*17*	*18*	$\sum$	$\frac{1}{N}\sum$	
Effectifs n_i	*3*	*1*	*1*	*1*	*3*	*1*			
Effectifs cumulés N_i	*3*	*4*	*5*	*6*	*9*	*N=10*			
$n_i x_i$	*36*	*14*	*15*	*16*	*51*	*18*	*150*	$\overline{X}_A = 15$	
$\lvert x_i - \overline{X}\rvert$	*3*	*1*	*0*	*1*	*2*	*3*			
$n_i \lvert x_i - \overline{X}\rvert$	*9*	*1*	*0*	*1*	*6*	*3*	*20*	$e_A = 2$	
$(x_i - \overline{X})^2$	*9*	*1*	*0*	*1*	*4*	*9*			
$n_i (x_i - \overline{X})^2$	*27*	*1*	*0*	*1*	*12*	*9*	*50*	$V_A = 5$	$\sigma_A = 2.23$

Pour l'étudiant B on a le tableau :

Notes x_i	*12*	*14*	*15*	*16*	*17*	*18*	$\sum$	$\frac{1}{N}\sum$	
Effectifs n_i	*0*	*2*	*6*	*2*	*0*	*0*			
Effectifs cumulés N_i	*0*	*2*	*8*	*10*	*10*	*N=10*			
$n_i x_i$	*0*	*28*	*90*	*32*	*0*	*0*	*150*	$\overline{X}_B = 15$	
$\lvert x_i - \overline{X}\rvert$	*3*	*1*	*0*	*1*	*2*	*3*			
$n_i \lvert x_i - \overline{X}\rvert$	*0*	*2*	*0*	*2*	*0*	*0*	*4*	$e_B = 0.4$	
$(x_i - \overline{X})^2$	*9*	*1*	*0*	*1*	*4*	*9*			
$n_i (x_i - \overline{X})^2$	*0*	*2*	*0*	*2*	*0*	*0*	*4*	$V_B = 0.4$	$\sigma_B = 0.63$

On voit que $\sigma_B < 1$ ce qui s'explique par le fait que les notes de l'étudiant B sont concentrées autour de la moyenne arithmétique. Ceci dit, le travail de l'étudiant B est régulier.

Exemple 1.15 *Reprenons l'exemple 1.10 :*

Notes I_i	*[0,5[*	*[5,8[*	*[8,12[*	*[12,15[*	*[15,20[*	$\sum$	$\frac{1}{N}\sum$	
Effectifs n_i	*10*	*8*	*12*	*11*	*9*			
Effectifs cumulés N_i	*10*	*18*	*30*	*41*	*N=50*			
Centres c_i	*2.5*	*6.5*	*10*	*13.5*	*17.5*			
$n_i c_i$	*25*	*52*	*120*	*148.5*	*157.5*	*503*	$\overline{X} = 10.06$	
$\lvert c_i - \overline{X}\rvert$	*7.56*	*3.56*	*0.06*	*3.44*	*7.44*			
$n_i \lvert c_i - \overline{X}\rvert$	*75.6*	*28.48*	*0.72*	*37.84*	*66.96*	*209.6*	$e = 4.192$	
$(c_i - \overline{X})^2$	*57.15*	*12.67*	*0.0036*	*11.83*	*55.35*			
$n_i (c_i - \overline{X})^2$	*571.5*	*101.36*	*0.0432*	*130.13*	*498.15*	*1301.18*	$V = 26.02$	$\sigma = 5.1$

1.3 Distributions statistiques à deux caractères

Les séries statistiques présentées au cours du premier paragraphe sont, le plus souvent, obtenues par dépouillement d'une enquête. Il est bien évident (pour de simples raisons économiques) que de nombreuses questions sont posées au cours d'une même enquête. Les différentes réponses sont regroupées sur une même fiche pour chaque individu sondé. Il est donc possible d'en extraire, non seulement la distribution des réponses à chaque question proposée, mais également la distribution des réponses à un couple (ou un triplet, ...) de questions. Le problème posé est alors de savoir s'il existe globalement une relation entre les réponses aux différentes questions. Le statisticien parle dans ce cas de corrélation entre les réponses. Le but de ce paragraphe est l'étude de cette notion de corrélation dans le cas de deux questions et, le cas échéant, de déterminer cette relation.

1.3.1 Définitions et exemples

Soit une population finie Ω dont chaque élément présente deux caractères X et Y.

Définition 1.7 *On appelle série statistique double sur Ω, toute application*

$$\begin{aligned} \overrightarrow{C} \ : \quad & \Omega \longrightarrow \mathbb{R}^2 \\ & w \longmapsto \overrightarrow{C}(w) := (X(w), Y(w)) \end{aligned}$$

qui associe à chaque individu w de Ω, le couple de modalités $(X(w), Y(w))$.
Par abus, on note $\overrightarrow{C} = (X, Y)$ la série statistique double.

Exemple 1.16 – *À chaque individu, on associe son poids (X) et sa taille (Y).*
– *À chaque voiture, on associe sa marque (X) et sa couleur (Y).*

La population Ω étant finie, il en est de même des ensembles $X(\Omega)$ et $Y(\Omega)$. Ainsi, nous écrivons $X(\Omega) = \{x_1, x_2, ..., x_n\}$ et $Y(\Omega) = \{y_1, y_2, ..., y_p\}$ en supposant que l'on a $x_1 < x_2 < ... < x_n$ et $y_1 < y_2 < ... < y_p$.

Définition 1.8 *Soit $\overrightarrow{C} = (X, Y)$ une série statistique double.*
– *On appelle effectif du couple (x_i, y_j) pour $(i, j) \in \{1, 2, ..., n\} \times \{1, 2, ..., p\}$ le nombre n_{ij} défini par*

$$n_{ij} = Card\Big(\Big\{w \in \Omega / \overrightarrow{C}(w) = (x_i, y_j)\Big\}\Big).$$

– $Card(\Omega) = \sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}$ *s'appelle* **l'effectif total** *de la série, nous le noterons N.*
– *On appelle effectif associé à la modalité x_i (resp. y_j) de la variable X (resp. Y) le nombre noté $n_{i.}$ (resp. $n_{.j}$) défini par*

$$n_{i\bullet} = \sum_{j=1}^{p} n_{ij} \qquad \Big(resp.\ n_{\bullet j} = \sum_{i=1}^{n} n_{ij}\Big).$$

– *La famille $(x_i, n_{i\bullet})_{i\in\{1,2,...,n\}}$ s'appelle* **première série marginale**.
– *La famille $(y_j, n_{\bullet j})_{j\in\{1,2,...,p\}}$ s'appelle* **deuxième série marginale**.
– *La famille $((x_i, y_j), n_{ij})_{(i,j)\in\{1,2,...,n\}\times\{1,2,...,p\}}$ s'appelle série statistique double discrète associée à la statistique $\overrightarrow{C}$.*
– *Pour $j \in \{1, 2, ..., p\}$ fixé, la famille $(x_i, n_{ij})_{i\in\{1,2,...,n\}}$ s'appelle* **série statistique conditionnelle** *de X sachant que le second caractère Y vaut y_j, on la note $X_{/Y=y_j}$.*
– *Pour $i \in \{1, 2, ..., n\}$ fixé, la famille $(y_j, n_{ij})_{j\in\{1,2,...,p\}}$ s'appelle* **série statistique conditionnelle** *de Y sachant que le premier caractère X vaut x_i, on la note $Y_{/X=x_i}$.*
– *Le nombre $f_{ij} = \frac{n_{ij}}{N}$ est* **appelé fréquence du couple** (x_i, y_j) *ou* **fréquence conjointe**, *de plus on a $\sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij} = 1$.*
– *Les nombres $f_{i\bullet} = \frac{n_{i\bullet}}{N}$ et $f_{\bullet j} = \frac{n_{\bullet j}}{N}$ sont appelés* **fréquences marginales**.

Remarque 1.7 *Il est facile de déterminer les séries statistiques marginales d'une série double. En fait $n_{i\bullet}$ n'est autre que la somme des éléments de $i^{ème}$ ligne du tableau représentant le couple. Il suffit donc d'ajouter une colonne à droite de ce tableau pour y placer ces sommes. De même $n_{\bullet j}$ s'obtient en faisant la somme des éléments de la $j^{ème}$ colonne, on ajoutera donc de la même façon une ligne au tableau. D'ou le nom de séries "marginales".*

Si l'effectif totale est grand et si l'un et/ou l'autre des caractères prend un grand nombre de valeurs, on peut être ramené à regrouper les valeurs de l'un et/ou de l'autre des caractères par intervalles. On peut définir ainsi des séries statistiques doubles regroupées ou semi-regroupées.
Si la population Ω comporte peu d'éléments, on présente généralement une statistique à deux caractères à l'aide d'un tableau à 3 lignes obtenu en plaçant en colonne l'individu w et les valeurs des caractères $X(w)$ et $Y(w)$.

Exemple 1.17 *La statistique suivante donne pour chaque pays les taux de chômage et d'inflation correspondant à l'année 2010 :*

pays	*Allemagne*	*Belgique*	*France*	*Italie*	*Luxemburg*	*Pays-Bas*
Chômage X	*7.1*	*8.7*	*9.8*	*8.7*	*5.8*	*4.2*
Inflation Y	*1.3*	*2.0*	*1.6*	*1.6*	*2.3*	*1.3*

La série statistique associé s'obtient simplement en oubliant la première ligne de ce tableau.

Par contre, si la population Ω a un effectif total assez grand, il est commode de présenter la série statistique double sous la forme d'un tableau à n lignes et p colonnes obtenu en plaçant l'effectif n_{ij} de (x_i, y_j) à l'intersection de la $i^{ème}$ ligne et de la $j^{ème}$ colonne de ce tableau.

Exemple 1.18 *On reprend l'exemple 1.3*

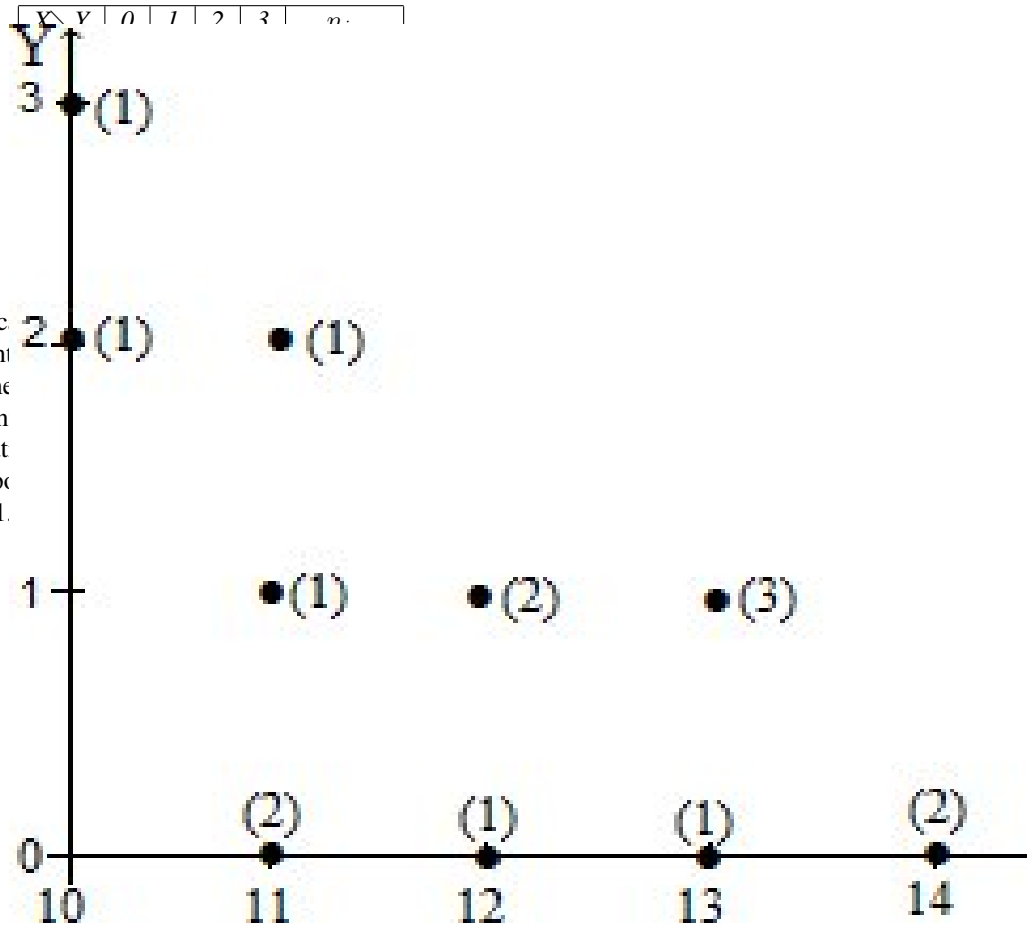


FIGURE 1.6 – Exemple 1.3.

On peut représenter une série statistique à c
le point $(x_i, y_j, 0)$ du plan horizontal et dont
statistique groupées, on construirai de même
à la fréquence correspondante. Ces représen
On préfère souvent représenter une série stat
à chaque point de coordonnée (x_i, y_j) le pc
rectangle par son centre). Ainsi, la figure 1.

sont pas représentés).

1.3.2 Distributions marginales

Les effectifs marginaux $n_{i\bullet}$ (ou $n_{\bullet j}$) définissent une distribution marginale selon le caractère X (ou Y) seul.

Caractéristiques marginales

La moyenne marginale en X est

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{n} n_{i\bullet} x_i = \sum_{i=1}^{n} f_{i\bullet} x_i.$$

La variance en X est

$$V(X) = \frac{1}{N}\sum_{i=1}^{n} n_{i\bullet}(x_i - \overline{X})^2 = \Big[\frac{1}{N}\sum_{i=1}^{n} n_{i\bullet} x_i^2\Big] - \overline{X}^2 = \sum_{i=1}^{n} f_{i\bullet} x_i^2 - \overline{X}^2.$$

L'écart-type est

$$\sigma_X = \sqrt{V(X)}.$$

La moyenne marginale en Y est

$$\overline{Y} = \frac{1}{N}\sum_{j=1}^{p} n_{\bullet j} y_j = \sum_{j=1}^{p} f_{\bullet j} y_j.$$

La variance en Y est

$$V(Y) = \frac{1}{N}\sum_{j=1}^{p} n_{\bullet j}(y_j - \overline{Y})^2 = \Big[\frac{1}{N}\sum_{j=1}^{p} n_{\bullet j} y_j^2\Big] - \overline{Y}^2 = \sum_{j=1}^{p} f_{\bullet j} y_j^2 - \overline{Y}^2.$$

L'écart-type est

$$\sigma_Y = \sqrt{V(Y)}.$$

Définition 1.9 *On appelle covariance du couple (X,Y) et on note $cov(X,Y)$ (ou σ_{xy}) la moyenne de $(X-\overline{X})(Y-\overline{Y})$ c'est-à-dire*

$$\begin{aligned} cov(X,Y) &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}(x_i-\overline{X})(y_j-\overline{Y}) \\ &= \Big(\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}x_i y_j\Big) - \overline{X}\,\overline{Y}. \end{aligned}$$

Propriété 1.1 *Soit (X,Y) une statistique double et soit a, b deux réels. Alors on a :*
- *$cov(X,Y) = cov(Y,X)$ (Symétrie de la covariance).*
- *$cov(a\cdot X + b, Y) = a\cdot cov(X,Y)$ (Linéarité par rapport à X).*
- *$cov(X, a\cdot Y + b) = a\cdot cov(X,Y)$ (Linéarité par rapport à Y).*

Remarque 1.8
- *La covariance peut prendre des valeurs positives, négatives ou nulles.*
- *Si $X = Y$ alors $cov(X,Y) = V(X)$.*

1.3.3 Fréquences conditionnelles, distributions conditionnelles, indépendance

Soit $((x_i,y_j),n_{ij})_{(i,j)\in\{1,2,\ldots,n\}\times\{1,2,\ldots,p\}}$ une série statistique double.

$$N = \sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij} = \sum_{i=1}^{n} n_{i\bullet} = \sum_{j=1}^{p} n_{\bullet j}, \text{ l'effectif total.}$$

$$f_{ij} = \frac{n_{ij}}{N}, \text{ la fréquence du couple } (x_i,y_j).$$

$$n_{i\bullet} = \sum_{j=1}^{p} n_{ij}, \text{ étant l'effectif marginal de } x_i.$$

$$n_{\bullet j} = \sum_{i=1}^{n} n_{ij}, \text{ étant l'effectif marginal de } y_j.$$

$$f_{i\bullet} = \frac{n_{i\bullet}}{N}, \text{ la fréquence marginale de } x_i.$$

$$f_{\bullet j} = \frac{n_{\bullet j}}{N}, \text{ la fréquence marginale de } y_j.$$

Définition 1.10
- *On appelle fréquence conditionnelle de x_i sachant y_j, le nombre*

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}.$$

- *On appelle fréquence conditionnelle de y_j sachant x_i, le nombre*

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}.$$

- *La distribution $(x_i, f_{i/j})_{i\in\{1,2,\ldots,n\}}$ est appelée la distribution conditionnelle des fréquences de X sachant que $Y = y_j$.*
- *La distribution $(y_j, f_{j/i})_{j\in\{1,2,\ldots,p\}}$ est appelée la distribution conditionnelle des fréquences de Y sachant que $X = x_i$.*

Ainsi, on remarque qu'on a :

$$f_{ij} = f_{i/j} \times f_{\bullet j} = f_{j/i} \times f_{i\bullet}.$$

Définition 1.11 *On dit que les caractères observés X et Y sont statistiquement indépendants si et seulement si*

$$\forall (i,j) \in \{1,2,...,n\} \times \{1,2,...,p\} : \ f_{ij} = f_{i\bullet} \times f_{\bullet j}.$$

Autrement dit, si :

$$f_{i/j} = f_{i\bullet} = f_{\bullet j}.$$

1.3.4 Coefficient de corrélation

Étudier la corrélation entre deux (ou plusieurs) variables statistiques, c'est étudier l'intensité de la liaison qui peut exister entre ces deux (ou plusieurs) variables. Lorsque le graphe montre une relation entre les mesures, on peut quantifier cette relation à l'aide d'un coefficient de corrélation. Si la liaison recherchée est une relation affine, on utilise le coefficient de corrélation linéaire.

Définition 1.12 *Soient X et Y deux caractères tels que $\sigma_X \neq 0$ et $\sigma_Y \neq 0$. On appelle coefficient de corrélation linéaire, le nombre réel*

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y}.$$

Proposition 1.4 *Si X et Y sont deux caractères tels que $\sigma_X \neq 0$ et $\sigma_Y \neq 0$, alors on a :*

$$-1 \leq \rho(X,Y) \leq 1.$$

Avant de donner la démonstration, nous rappelons l'inégalité suivante :

Théorème 1.1 *Inégalité de Cauchy-Schwarz*
Si $a_1, a_2, ..., a_n$ et $b_1, b_2, ..., b_n$ sont des nombres réels, alors on a

$$\Big(\sum_{i=1}^{n} a_i b_i\Big)^2 \leq \Big(\sum_{i=1}^{n} a_i^2\Big)\Big(\sum_{i=1}^{n} b_i^2\Big).$$

Preuve.(de la proposition 1.3) En appliquant l'inégalité de Cauchy-Schwartz à double reprise, on a

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}(x_i - \overline{X})(y_j - \overline{Y}) &= \sum_{i=1}^{n}\sum_{j=1}^{p} \sqrt{n_{ij}}(x_i - \overline{X})\sqrt{n_{ij}}(y_j - \overline{Y}) \\
&\leq \sum_{i=1}^{n}\Big[\sum_{j=1}^{p} n_{ij}(x_i - \overline{X})^2\Big]^{\frac{1}{2}}\Big[\sum_{j=1}^{p} n_{ij}(y_j - \overline{Y})^2\Big]^{\frac{1}{2}} \\
&\leq \Big[\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}(x_i - \overline{X})^2\Big]^{\frac{1}{2}}\Big[\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}(y_j - \overline{Y})^2\Big]^{\frac{1}{2}} \\
&= \Big[\sum_{i=1}^{n} n_{i\bullet}(x_i - \overline{X})^2\Big]^{\frac{1}{2}}\Big[\sum_{j=1}^{p} n_{\bullet j}(y_j - \overline{Y})^2\Big]^{\frac{1}{2}}.
\end{aligned}$$

Par conséquent, $\rho(X,Y)^2 - 1 = \dfrac{\Big[\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}(x_i - \overline{X})(y_j - \overline{Y})\Big]^2 - \Big(\sum_{i=1}^{n} n_{i\bullet}(x_i - \overline{X})^2\Big)\Big(\sum_{j=1}^{p} n_{\bullet j}(y_j - \overline{Y})^2\Big)}{\Big(\sum_{i=1}^{n} n_{i\bullet}(x_i - \overline{X})^2\Big)\Big(\sum_{j=1}^{p} n_{\bullet j}(y_j - \overline{Y})^2\Big)} \leq 0.$

D'où

$$|\rho(X,Y)| \leq 1.$$

Remarque 1.9 *L'égalité $\rho(X,Y) = \pm 1$ a eu lieu si et seulement si il existe $a \in \mathbb{R} \setminus \{0\}$, et $b \in \mathbb{R}$ tels que $Y = a \cdot X + b$.*

Si $\rho(X,Y)=1$ alors l'une des variables est fonction affine **croissante** de l'autre variable et si $\rho(X,Y)=-1$ alors la fonction affine est **décroissante**. Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte. On emploie l'expression **fortement corrélées** pour qualifier une telle corrélation.
Une corrélation égale à 0 signifie que les variables sont **linéairement indépendant**.

Exemple 1.19 *Soit* $((x_i,y_i),1)_{i\in\{1,2,...,n\}}$ *une série statistique double.*
- *Si* $\forall i\in\{1,2,...,n\}: x_i=3y_i$, *alors*

$$\begin{aligned}
\overline{X} &= \frac{1}{N}\sum_{i=1}^{n}x_i=\frac{1}{N}\sum_{i=1}^{n}3y_i=3\overline{Y}.\\
V(X) &= \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{X})^2=9\frac{1}{N}\sum_{i=1}^{n}(y_i-\overline{Y})^2=9V(Y).\\
cov(X,Y) &= \frac{1}{N}\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})=3\frac{1}{N}\sum_{i=1}^{n}(y_i-\overline{Y})^2=3V(Y).\\
\rho(X,Y) &= \frac{cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}=\frac{3V(Y)}{\sqrt{9V(Y)}\sqrt{V(Y)}}=1.
\end{aligned}$$

- *Si* $\forall i\in\{1,2,...,n\}: x_i=-2y_i$, *alors de la même manière on trouve*

$$\rho(X,Y)=-1.$$

Exercice 1.1 *Déterminer le coefficient de corrélation de la série statistique* $(10,30)$, $(20,60)$, $(30,90)$, $(40,120)$, $(50,150)$, $(60,180)$.

Solution: L'effectif total est $N=6$.

$$\begin{aligned}
\overline{X} &= \frac{1}{6}(10+20+30+40+50+60)=35.\\
\overline{Y} &= \frac{1}{6}(30+60+90+120+150+180)=105.\\
V(X) &= \frac{1}{6}((10-35)^2+(20-35)^2+(30-35)^2+(40-35)^2+(50-35)^2+(60-35)^2)\\
&= \frac{1}{6}(625+225+25+25+225+625)=291.66.\\
\sigma_X &= \sqrt{V(X)}\simeq 17.07.\\
V(Y) &= \frac{1}{6}((30-105)^2+(60-105)^2+(90-105)^2+(120-105)^2+(150-105)^2\\
&\quad +(180-105)^2)\\
&= \frac{1}{6}(5625+2025+225+225+2025+5625)=2625.
\end{aligned}$$

$$\begin{aligned}
\sigma_Y &= \sqrt{V(Y)}\simeq 51.23.\\
cov(X,Y) &= \frac{1}{6}\sum(x_i-\overline{X})(y_i-\overline{Y})\\
&= \frac{1}{6}((10-35)(30-105)+(20-35)(60-105)+(30-35)(90-105)\\
&\quad +(40-35)(120-105)+(50-35)(150-105)+(60-35)(180-105))\\
&= \frac{1}{6}(1875+675+75+75+675+1875)\\
&= 875.\\
\rho(X,Y) &= \frac{cov(X,Y)}{\sigma_X\times\sigma_Y}=\frac{875}{17.07\times 51.23}=1.
\end{aligned}$$

On voit que le coefficient de corrélation est 1, ce qui explique que les variables sont fortement corrélées.

Remarque 1.10 *Le coefficient de corrélation n'est pas sensible aux unités de chacune des variables. Ainsi, par exemple le coefficient de corrélation linéaire entre l'âge et le poids d'un individu sera identique que l'âge soit mesuré en semaine, en mois ou en année(s). Cependant, il est très sensible à la présence des valeurs aberrantes et/ou extrêmes dans l'ensemble des données (valeurs très éloignées de la majorité des autres, pouvant être considérées comme des exceptions).*

Remarque 1.11 *Il ne faut pas croire qu'un coefficient de corrélation élevé induit une relation de causalité entre les deux caractères mesurés. En réalité, les deux caractères peuvent être corrélés à un même phénomène-source : un troisième caractère non mesuré et dont dépendent les deux autres. Par exemple : le nombre de coups de soleil observés dans une station balnéaire est fortement corrélé au nombre de lunette de soleil vendues, mais aucun des deux caractères n'est bien sûr la cause de l'autre.*

CHAPITRE 2

Ajustement linéaire

2.1 Introduction

Dans le domaine des sciences appliquées, on observe fréquemment des phénomènes tels qu'il est possible de supposer l'existence d'une liaison entre deux variables. Par exemple :
- les dépenses annuelles d'un ménage sont fonction des revenus de la famille.
- la durée de vie d'une ampoule électrique peut être liée à son rendement énergétique.

Supposons qu'on a une population de taille (effectif total) N, telle que pour chaque individu on observe deux caractères différents X et Y. Le problème qui se pose consiste à chercher une relation entre les deux caractères observés X et Y. A chaque élément i, $1 \leq i \leq N$, de la population on associe un couple de valeurs (x_i, y_i) qu'on représente graphiquement par un point $M_i(x_i, y_i)$ du plan. Ainsi, on obtient un nuage de points qui constitue ce qu'on appelle le **diagramme de dispersion**.
Ajuster un ensemble de points, c'est déterminer une courbe simple $(\mathcal{C})$ aussi proche que possible des points $M_i(x_i, y_i)$. La courbe $(\mathcal{C})$ peut être une parabole, une droite ou peut représentée une fonction exponentielle. On parle d'ajustement linéaire lorsque la courbe $(\mathcal{C})$ est une droite.

Définition 2.1 *On dit qu'il y a corrélation linéaire entre deux variables observées sur les éléments d'une population si les variations de ces deux variables se produisent dans le même sens (corrélation positive) ou en sens contraires (corrélation négative).*

2.2 Alignement statistique

Soit $\overrightarrow{C} : \Omega \to \mathbb{R}^2$ $w \to \overrightarrow{C}(w) = (X(w), Y(w))$ une statistique double. Nous supposons d'abord dans ce paragraphe que $\overrightarrow{C}$ soit une application injective, c'est-à-dire que les couples de valeurs des caractères associés à deux individus distincts soient distincts. On peut donc représenter $\overrightarrow{C} = (X, Y)$ par la série statistique $((x_i, y_i), 1)_{i\in\{1,2,\ldots,n\}}$ observés sur une population de taille n. La moyenne en X est

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La variance en X est

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \Big[\frac{1}{n}\sum_{i=1}^{n} x_i^2\Big] - \overline{X}^2.$$

L'écart-type est

$$\sigma_X = \sqrt{V(X)}.$$

La moyenne en Y est

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

La variance en Y est

$$V(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2 = \Big[\frac{1}{n}\sum_{i=1}^{n} y_i^2\Big] - \overline{Y}^2.$$

L'écart-type est

$$\sigma_Y = \sqrt{V(Y)}.$$

La covariance est

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\,\overline{Y}.$$

Définition 2.2 *On appelle coefficient de corrélation de la série statistique double* $((x_i, y_i), 1)_{i\in\{1,2,\dots,n\}}$ *le nombre réel*

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y} = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(y_i - \overline{Y})^2}}.$$

Remarque 2.1 $-1 \leq \rho(X,Y) \leq 1$.

Le coefficient de corrélation linéaire ρ
- Si le coefficient de corrélation e
sens). Si $\rho(X,Y) = 1$ alors la c
- Si le coefficient de corrélation e
sens contraires). Si $\rho(X,Y) =$
- Pour éviter les problèmes de sig
corrélation définit dans la défini
- Si les variables X et Y sont ind
- Si le coefficient de corrélation
indépendantes). On peut cepen
Gaussiennes).

Le coefficient de corrélation linéaire
peuvent être corrélées sans que les va
variations des deux variables sont due
une augmentation de la vente des crèm
produits. Il faut toujours être très prud

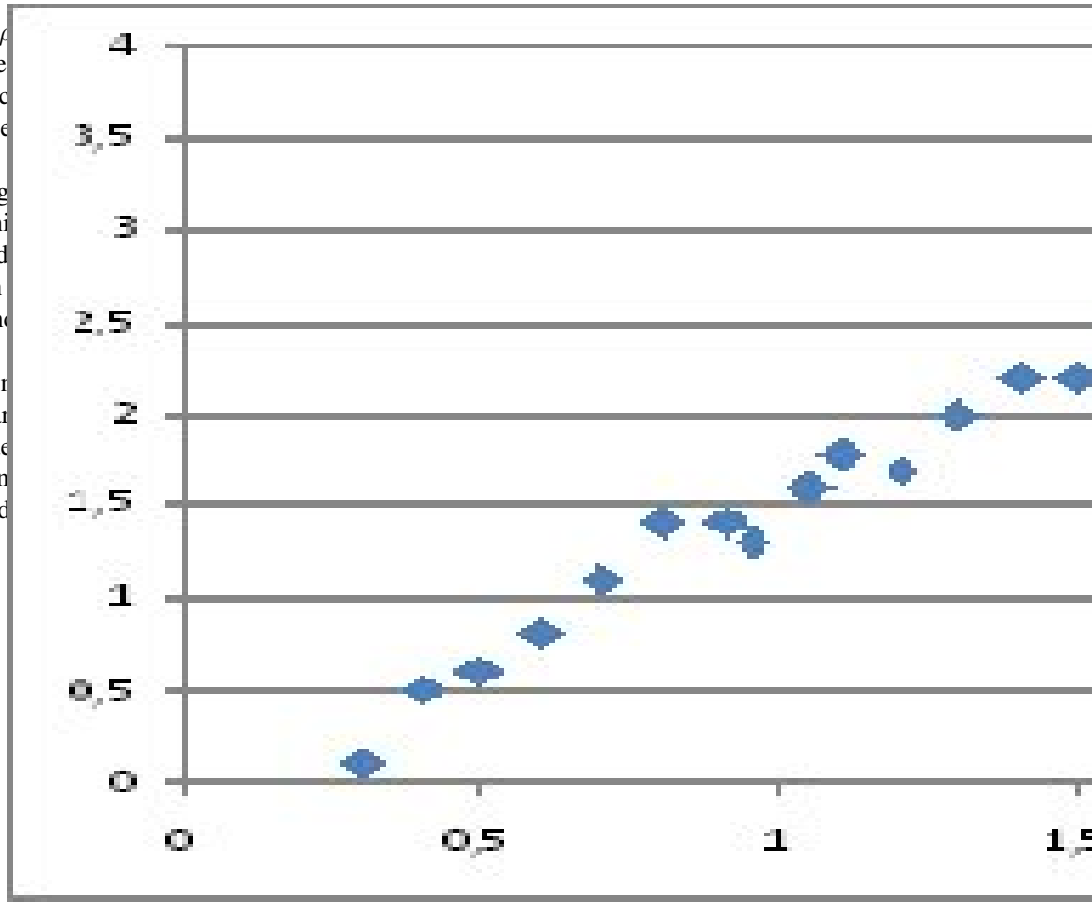


FIGURE 2.1 – Forte corrélation "positive."

1. Karl Pearson, Anglais, (1857-1936).

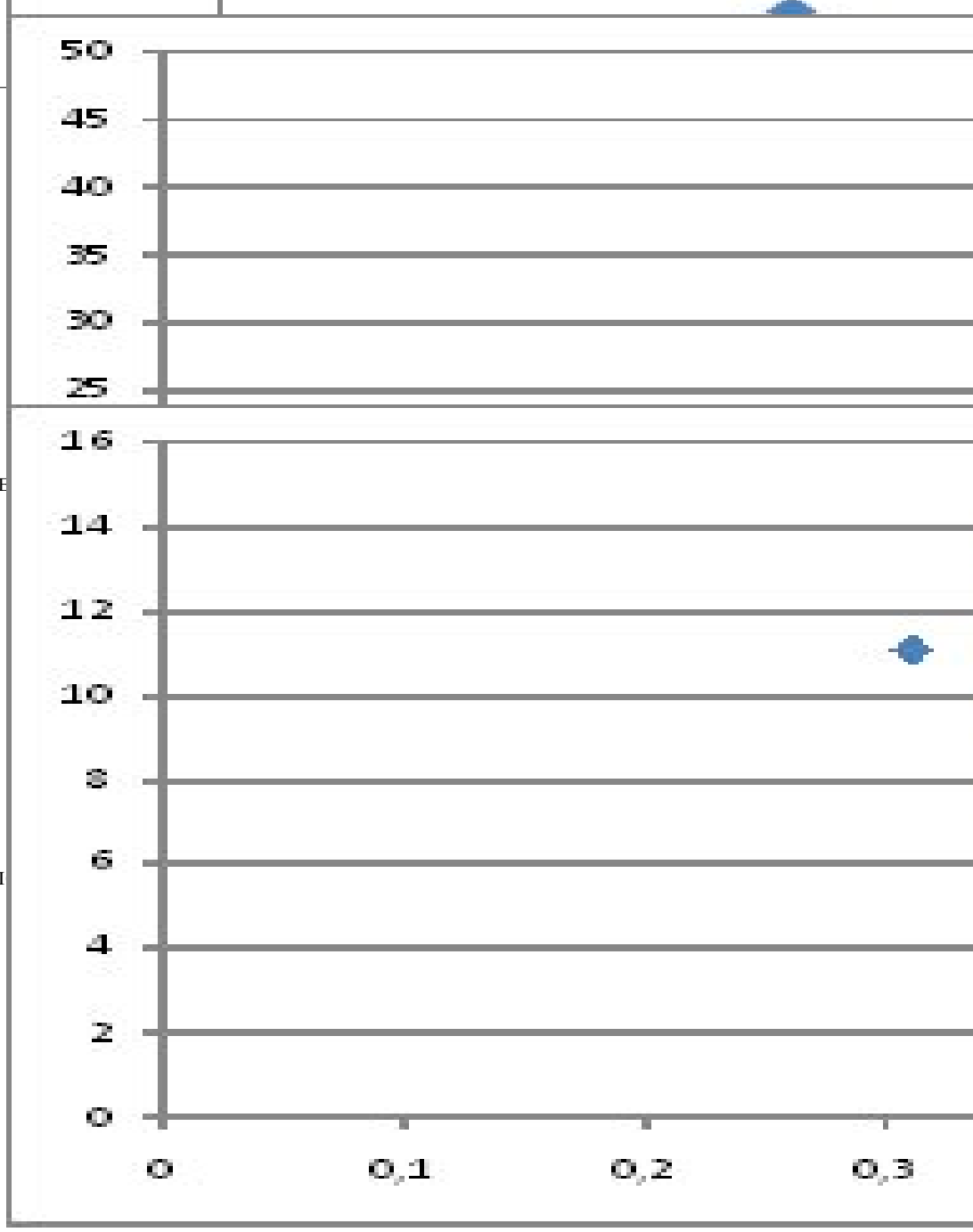


FIGURE 2.4 – Pas de corrélation : Variables indépendantes.

Remarque 2.2 *Importante.*
Il convient de prêter une attention toute particulière aux unités choisies pour construire le nuage. En effect, une unité trop petite sur l'un et/ou sur l'autre des axes écrase ce nuage et peut laisser croire à un alignement qui n'a pas de sens statistique. Il faut donc faire en sorte que celui-ci remplisse au mieux la figure, quitte pour cela à effectuer sur l'un et/ou sur l'autre des axes un changement d'origine et/ou d'unité.

2.2.1 Méthode des moindres carrés

La méthode graphique d'ajustement est bien évidemment empirique et subjective. On se limite au cas de l'ajustement linéaire, c'est-à-dire lorsque les deux caractères X et Y observés sur une même population semblent être liés par une droite. Même dans le cas d'un phénomène linéaire, une détermination graphique de la droite d'ajustement (droite approximant "au mieux" le nuage de points) mènera à des résultats différents selon les opérateurs. Il est donc indispensable de définir rigoureusement la notion de courbe d'ajustement, de telle sorte qu'il n'y ait qu'une seule réponse possible.
Soit $M_i(x_i, y_i)$, $i \in \{1, 2, ..., n\}$, un nuage de points et soit (D) la droite d'équation $y = ax + b$. Notons H_i la projection de

M_i sur la droite (D) parallèlement à l'axe (Oy). Posons :

Nous allons montrer qu'il existe une u

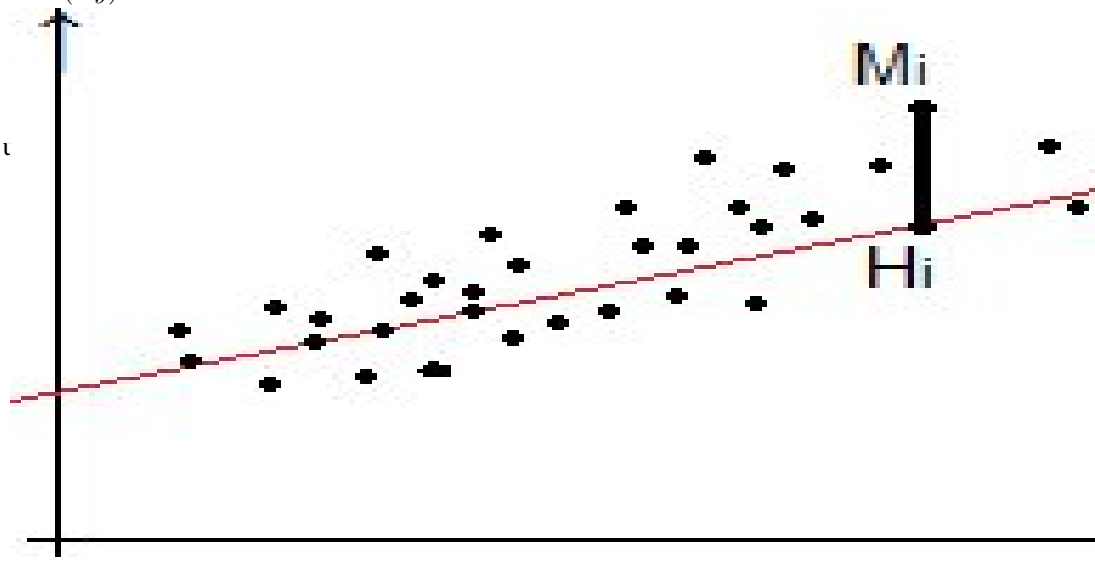


FIGURE 2.5 –

Définition 2.3 *On appelle* **droite des moindres carrés** *de Y en X, ou également* **droite de régression** *de Y en X, l'unique droite rendant minimale la somme $\delta(a,b)$.*

Remarque 2.3 *A ne pas croire qu'il s'agit de la somme des carrés des distances des points M_i à la droite (D).*

Comme le point M_i a pour coordonnées (x_i, y_i), le point H_i a pour coordonnées $(x_i, ax_i + b)$. Par conséquent :

$$\delta(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Ainsi, il s'agit de trouver les deux réels a et b tels que la somme $\delta(a,b)$ soit minimale. Pour cela, nous donnons deux méthodes.
$1^{\text{ère}}$ **méthode : Utilisation de la dérivée.**
Il suffit d'annuler les dérivées partielles de $\delta(a,b)$ par rapport à a et à b.

$$\begin{cases} \dfrac{\partial \delta}{\partial a}(a,b) = \displaystyle\sum_{i=1}^{n} (-2x_i)(y_i - ax_i - b) = 0, \\ \dfrac{\partial \delta}{\partial b}(a,b) = \displaystyle\sum_{i=1}^{n} (-2)(y_i - ax_i - b) = 0, \end{cases}$$

$$\begin{cases} \displaystyle\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0, \\ \displaystyle\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - bn = 0, \end{cases}$$

$$\begin{cases} \displaystyle\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - bn\overline{X} = 0, \\ \overline{Y} - a\overline{X} - b = 0. \end{cases}$$

En multipliant les deux termes de la $1^{\text{ère}}$ équation par $\frac{1}{n}$ et les deux termes de la $2^{\text{ème}}$ équation par $-\overline{X}$, nous obtenons le système suivant :

$$\begin{cases} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i y_i - a \frac{1}{n} \sum_{i=1}^{n} x_i^2 - b\overline{X} = 0, \quad (1) \\ -\overline{X}\,\overline{Y} + a\overline{X}^2 + b\overline{X} = 0. \qquad (2) \end{cases}$$

En faisant la somme terme à terme des deux équations (1) et (2), on obtient

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - a\Big(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}\Big) = 0.$$

Soit

$$cov(X,Y) - aV(X) = 0.$$

Ainsi,

$$a = \frac{cov(X,Y)}{V(X)}.$$

De l'équation (2) on tire

$$b = \overline{Y} - a\overline{X}.$$

$2^{\text{ème}}$ **méthode : technique du trinôme.**
Nous allons écrire $\delta(a,b)$ sous la forme d'une somme de carrés, en l'écrivant sous forme d'un trinôme en b puis en le mettant sous sa forme canonique, ensuite sous forme d'un trinôme en a et en le mettant sous sa forme canonique tout en faisant apparaître les paramètres de position et de dispersion.

$$\begin{aligned}
\delta(a,b) &= \sum_{i=1}^{n}(y_i - ax_i - b)^2 \\
&= \sum_{i=1}^{n}\Big[b^2 - 2b(y_i - ax_i) + (y_i - ax_i)^2\Big] \\
&= nb^2 - 2b\sum_{i=1}^{n}(y_i - ax_i) + \sum_{i=1}^{n}(y_i - ax_i)^2 \\
&= nb^2 - 2nb(\overline{Y} - a\overline{X}) + \sum_{i=1}^{n}\Big(y_i^2 - 2ax_iy_i + a^2x_i^2\Big) \\
&= n(b - (\overline{Y} - a\overline{X}))^2 - n\Big(\overline{Y} - a\overline{X}\Big)^2 + \sum_{i=1}^{n}\Big(y_i^2 - 2ax_iy_i + a^2x_i^2\Big) \\
&= n\Big(b - (\overline{Y} - a\overline{X})\Big)^2 - n\overline{Y}^2 + 2na\overline{X}\,\overline{Y} - na^2\overline{X}^2 + \sum_{i=1}^{n}y_i^2 - 2a\sum_{i=1}^{n}x_iy_i + a^2\sum_{i=1}^{n}x_i^2 \\
&= n\Big(b - (\overline{Y} - a\overline{X})\Big)^2 + na^2\Big(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{X}^2\Big) - 2na\Big(\frac{1}{n}\sum_{i=1}^{n}x_iy_i - \overline{X}\,\overline{Y}\Big) + n\Big(\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \overline{Y}^2\Big) \\
&= n\Big[\Big(b - (\overline{Y} - a\overline{X})\Big)^2 + a^2V(X) - 2a\,cov(X,Y) + V(Y)\Big] \\
&= n\Big[\Big(b - (\overline{Y} - a\overline{X})\Big)^2 + \Big(a\sqrt{V(X)} - \frac{cov(X,Y)}{\sqrt{V(X)}}\Big)^2 + \frac{V(X)V(Y) - cov(X,Y)^2}{V(X)}\Big].
\end{aligned}$$

Ainsi, la somme $\delta(a,b)$ est minimale si et seulement si

$$a = \frac{cov(X,Y)}{V(X)}$$

et

$$b = \overline{Y} - a\overline{X}.$$

Proposition 2.1 *La droite des moindres carrés (de régression) de Y en X, notée $(D_{Y/X})$, est la droite d'équation $y = ax + b$ avec*

$$a = \frac{cov(X,Y)}{V(X)}$$

et

$$b = \overline{Y} - a\overline{X}.$$

Cette dernière relation exprime le fait que la droite d'ajustement passe par le point de coordonnées $(\overline{X},\overline{Y})$, c'est-à-dire par le barycentre (le point moyen) des points du nuage.

Remarque 2.4 *La somme $\delta(a,b)$ étant positive, ainsi le minimum*

$$\min_{(a,b)\in\mathbb{R}^2} \delta(a,b) = n\frac{V(X)V(Y) - cov(X,Y)^2}{V(X)} \geq 0.$$

Par conséquent, on retrouve l'inégalité $V(X)V(Y) - cov(X,Y)^2 \geq 0$, c'est-à-dire $\rho(X,Y)^2 \leq 1$.
L'égalité a eu lieu si et seulement si les points sont alignés.

On admet généralement qu'un coefficient de corrélation $|\rho(X,Y)| \geq 0.75$ justifie la recherche d'un alignement statistique, en l'absence de renseignements complémentaires.

Remarque 2.5 *On remarque que la droite des moindres carrés de Y en X a aussi pour équation :*

$$(D_{Y/X}): \quad y - \overline{Y} = a(x - \overline{X}).$$

On définit de même la droite de régression de X en Y notée par $(D_{X/Y})$.

$$(D_{X/Y}): \quad x - \overline{X} = \alpha(y - \overline{Y}) \quad \text{où } \alpha = \frac{cov(X,Y)}{V(Y)}.$$

La droite $(D_{X/Y})$ passe aussi par le point de coordonnées $(\overline{X}, \overline{Y})$.

2.2.2 Généralisation

Les calculs et les résultats précédents se généralisent aisément au cas où la statistique double $\overrightarrow{C}$ n'est pas injective. Ainsi, on regroupe les individus ayant le même couple de valeurs du caractère (x_i, y_j), $(i,j) \in \{1,2,...,n\}\times\{1,2,...,p\}$. Par suite les paramètres précédents deviennent :

$$\begin{aligned}
\overline{X} &= \frac{1}{N}\sum_{i=1}^{n} n_{i\bullet} x_i. \\
V(X) &= \frac{1}{N}\sum_{i=1}^{n} n_{i\bullet}(x_i - \overline{X})^2 = \Big[\frac{1}{N}\sum_{i=1}^{n} n_{i\bullet} x_i^2\Big] - \overline{X}^2. \\
\overline{Y} &= \frac{1}{N}\sum_{j=1}^{p} n_{\bullet j} y_j. \\
V(Y) &= \frac{1}{N}\sum_{j=1}^{p} n_{\bullet j}(y_j - \overline{Y})^2 = \Big[\frac{1}{N}\sum_{j=1}^{p} n_{\bullet j} y_j^2\Big] - \overline{Y}^2. \\
cov(X,Y) &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij}(x_i - \overline{X})(y_j - \overline{Y}) \\
&= \Big(\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij} x_i y_j\Big) - \overline{X}\,\overline{Y}.
\end{aligned}$$

Définition 2.4 *- La droite d'équation $(D_{Y/X})$: $y = ax + b$ avec $a = \dfrac{cov(X,Y)}{V(X)}$ et $b = \overline{Y} - a\overline{X}$, s'appelle* **droite de régression** *de Y en X.*
- La droite d'équation $(D_{X/Y})$: $x = \alpha y + \beta$ avec $\alpha = \dfrac{cov(X,Y)}{V(Y)}$ et $\beta = \overline{X} - \alpha\overline{Y}$, s'appelle **droite de régression** *de X en Y.*

Remarque 2.6 *Les deux droites $(D_{Y/X})$ et $(D_{X/Y})$ passent par le même point de coordonnées $(\overline{X}, \overline{Y})$.*

Exemple 2.1 *Soit à chercher l'équation de la droite de régression de Y en X associée à la série statistique suivante :* $(1,1), (2,12), (3,75), (4,145), (5,199), (6,256), (7,343)$.

Solution:

$X \setminus Y$	1	12	75	145	199	256	343	$n_{i\bullet}$	$n_{i\bullet}x_i$	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
1	1	0	0	0	0	0	0	1	1	-3	9
2	0	1	0	0	0	0	0	1	2	-2	4
3	0	0	1	0	0	0	0	1	3	-1	1
4	0	0	0	1	0	0	0	1	4	0	0
5	0	0	0	0	1	0	0	1	5	1	1
6	0	0	0	0	0	1	0	1	6	2	4
7	0	0	0	0	0	0	1	1	7	3	9
$n_{\bullet i}$	1	1	1	1	1	1	1	$N = 7$	$\overline{X} = 4$		$\sum = 28$
$n_{\bullet i}y_i$	1	12	75	145	199	256	343	$\overline{Y} = 147.286$			$V(X) = 4$
$y_i - \overline{Y}$	-146.286	-135.286	-72.286	-2.286	51.714	108.714	195.714				$\sigma_X = 2$
$(x_i - \overline{X})(y_i - \overline{Y})$	438.857	270.571	72.286	0	51.714	217.429	587.143	$\sum = 1638$	$cov(X,Y) = 234$		

Par suite

$$a = \frac{cov(X,Y)}{V(X)} = \frac{234}{4} = 58.5 \text{ et } b = \overline{Y} - a\overline{X} = 147.286 - 234 = -86.714.$$

Ainsi, la droite de régression de Y en X a pour équation :

$$(D_{Y/X}): \quad y = 58.5x - 86.714.$$

Exemple 2.2 *La répartition de* 100 *étudiants selon leurs résultats en Mathématiques (X) et en Biologie (Y) à donnée les résultats suivants :*

$X \setminus Y$	*7*	*11*	*12*	*15*
2	*9*	*2*	*1*	*0*
6	*7*	*27*	*4*	*1*
8	*1*	*3*	*15*	*4*
12	*0*	*0*	*4*	*17*
14	*0*	*1*	*2*	*2*

On se propose de déterminer l'équation de la droite de régression de Y en X.

Solution:

$X \setminus Y$	7	11	12	15	$n_{i\bullet}$	$n_{i\bullet}x_i$	$x_i - \overline{X}$	$(x_i - \overline{X})^2$	$n_{i\bullet}(x_i - \overline{X})^2$
2	9	2	1	0	12	108	-6.48	41.99	503.88
6	7	27	4	1	39	234	-2.48	6.15	239.85
8	1	3	15	4	23	184	-0.48	0.23	5.29
12	0	0	4	17	21	252	3.52	10.24	215.04
14	0	1	2	2	5	70	5.52	30.47	152.35
$n_{\bullet j}$	17	33	26	24	$N = 100$	$\overline{X} = 8.48$			$\sum = 1116.41$
$n_{\bullet j}y_j$	119	363	312	360	$\overline{Y} = 11.54$				
$y_j - \overline{Y}$	-4.54	-0.54	0.46	3.46					
$(y_j - \overline{Y})^2$	20.61	0.29	0.21	11.97					
$n_{\bullet j}(y_j - \overline{Y})^2$	350.37	9.57	5.46	287.28	$\sum = 652.68$				
$\sum_{i=1}^{5} n_{ij}(x_i - \overline{X})$	-76.16	-75.84	1.52	66.48					
$(y_j - \overline{Y}) \sum_{i=1}^{5} n_{ij}(x_i - \overline{X})$	345.76	40.95	0.69	230.02	$\sum = 617.42$				

- Moyennes marginales : $\overline{X} = 8.48$, $\overline{Y} = 11.54$.
- Variances : $V(X) = 11.1641$, $V(Y) = 6.5268$.
- Écart-types : $\sigma_X = \sqrt{V(X)} = 3.34$, $\sigma_Y = \sqrt{V(Y)} = 2.55$.
- Covariance : $cov(X,Y) = 6.1742$.

- Coefficient de corrélation linéaire : $\rho(X,Y) = \dfrac{cov(X,Y)}{\sigma_X \times \sigma_Y} \simeq 0.72$.
- Droite de régression de Y en X : $a = \dfrac{cov(X,Y)}{V(X)} \simeq 0.55$, $b \simeq 6.876$. Ainsi, l'équation de la droite est :

$$(D_{Y/X}) : \quad y = 0.55x + 6.876.$$

2.2.3 Exercices

Exercice 2.1 *On cherche s'il existe une relation entre la température et le nombre de glaces vendues. Les informations sont données par le tableau suivant :*

Température (Celcius)	*21*	*17*	*24*	*25*	*13*
Nbr de glaces vendues	*25*	*20*	*30*	*35*	*10*

1. *Calculer le coefficient de corrélation. Peut-on penser qu'il existe une relation entre les deux variables ?*
2. *Calculer l'équation de la droite de régression.*
3. *Quel serait alors le nombre de glaces vendues s'il faisait* 30 *degrés ?*
4. *Pour quelle température vendrait-on* 40 *glaces ?*

Solution: Soit Y le nombre de glaces vendues au moment où règne une température X. Les données ont été collectées en cinq prises. L'effectif total est donc $N = 5$. Ainsi, on a le tableau suivant :

Température (Celcius) X	21	17	24	25	13	$\overline{X} = 20$		
Nbr de glaces vendues Y	25	20	30	35	10	$\overline{Y} = 24$		
$x_i - \overline{X}$	1	-3	4	5	-7			
$(x_i - \overline{X})^2$	1	9	16	25	49	$\sum = 100$	$V(X) = 20$	$\sigma_X \simeq 4.47$
$y_i - \overline{Y}$	1	-4	6	11	-14			
$(y_i - \overline{Y})^2$	1	16	36	121	196	$\sum = 370$	$V(Y) = 74$	$\sigma_Y \simeq 8.6$
$(x_i - \overline{X})(y_i - \overline{Y})$	1	12	24	55	98	$\sum = 190$	$cov(X,Y) = 38$	

1. Le coefficient de corrélation :
$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y} = \frac{38}{8.6 \times 4.47} \simeq 0.988.$$
La valeur de $\rho(X,Y)$ montre que les variables sont fortement corrélées.
2. La droite de régression $(D_{Y/X})$ qui lie Y à X a pour équation $(D_{Y/X}) : \quad Y = aX + b$, avec
$$a = \frac{cov(X,Y)}{V(X)} = \frac{38}{20} = 1.9$$
et
$$b = \overline{Y} - a\overline{X} = -14.$$
Ainsi, $(D_{Y/X}) : \quad Y = 1.9X - 14$.
3. S'il faisait 30 degrés, on vendrait environ $Y = 1.9 \times 30 - 14 = 43$ glaces.
4. Si on vendait 40 glaces, la température serait d'environ $X = \frac{40+14}{1.9} \simeq 28.4$ degrés.

Exercice 2.2 *Un test de freinage a été effectué à partir de* 7 *voitures. Les résultats de ce test sont donnés par le tableau suivant :*

Vitesse (km/h)	*30*	*33*	*40*	*49*	*60*	*70*	*93*
Distance (m)	*6.50*	*5.30*	*14.45*	*11.23*	*20.21*	*38.45*	*50.42*

Le but de cet exercice est de déterminer la distance nécessaire à l'arrêt d'une voiture lancée à 100 *km/h.*

1. *Calculer le coefficient de corrélation. Peut-on penser qu'il existe une relation entre la vitesse et la distance de freinage d'une voiture ?*
2. *En utilisant la méthode des moindres carrés, proposez une équation de la droite de régression reliant la distance à la vitesse (distance=a×vitesse+b).*

3. *Quelle serait alors la distance nécessaire à l'arrêt d'une voiture lancée à* 100 *km/h ?*
4. *En utilisant la méthode des moindres carrés, proposez une équation de la droite de régression reliant la racine carrée de la distance à la vitesse (*$\sqrt{distance} = a \times$*vitesse+b).*
5. *Quelle serait alors la distance nécessaire à l'arrêt d'une voiture lancée à* 100 *km/h ?*
6. *Utilisez le coefficient de corrélation pour choisir l'une des deux modélisations.*
7. *Avec chacune des deux méthodes, quelle serait la vitesse d'une voiture qui s'arrête en* 40 *mètres.*

Solution: Notons la variable vitesse par V et la variable distance de freinage par D. On a le tableau suivant :

$Vitesse(km/h) v_i$	30	33	40	49	60	70	93	$\overline{V} = 53.57$
$Distance(m) d_i$	6.50	5.30	14.45	11.23	20.21	38.45	50.42	$\overline{D} = 20.94$
$v_i - \overline{V}$	-23.57	-20.57	-13.57	-4.57	6.43	16.43	39.43	
$(v_i - \overline{V})^2$	555.54	423.12	184.14	20.88	41.34	269.94	1554.72	$\sum = 3049.68$
$d_i - \overline{D}$	-14.44	-15.64	-6.49	-9.71	-0.73	17.51	29.48	
$(d_i - \overline{D})^2$	208.51	244.60	42.12	94.28	0.53	306.6	869.07	$\sum = 1765.71$
$(v_i - \overline{V})(d_i - \overline{D})$	340.35	321.71	88.06	44.37	-4.69	287.68	1162.39	$\sum = 2239.87$

Par conséquent, on a

- $V(V) = 435.66$ et $\sigma_V = 20.87$.
- $V(D) = 252.24$ et $\sigma_D = 15.88$.
- $cov(V, D) = 319.98$.

1. Ainsi, le coefficient de corrélation est
$$\rho(V, D) = \frac{cov(V, D)}{\sigma_V \times \sigma_D} = \frac{319.98}{20.87 \times 15.88} \simeq 0.965.$$
Nous pouvons déduire que les deux variables sont fortement corrélées.
2. Droite de régression de D en V.
$$a = \frac{cov(V, D)}{V(V)} = \frac{319.98}{435.66} \simeq 0.73 \quad \text{et;} \quad b = \overline{D} - a\overline{V} = 20.94 - 0.73 \times 53.57 \simeq -18.16.$$
Ainsi, l'équation la droite de régression est $\boxed{(D_{D/V}) : \quad D = 0.73 \times V - 18.16.}$
3. Une voiture roulant à 100 km/h s'arrêtera au bout de $D = 0.73 \times 100 - 18.16 = 54.84$ mètres.
4.

$Vitesse(km/h) v_i$	30	33	40	49	60	70	93	$\overline{V} = 53.57$
$Distance(m) \sqrt{d_i}$	2.55	2.30	3.80	3.35	4.50	6.20	7.10	$\overline{\sqrt{D}} = 4.26$
$v_i - \overline{V}$	-23.57	-20.57	-13.57	-4.57	6.43	16.43	39.43	
$(v_i - \overline{V})^2$	555.54	423.12	184.14	20.88	41.34	269.94	1554.72	$\sum = 3049.68$
$\sqrt{d_i} - \overline{\sqrt{D}}$	-1.71	-1.96	-0.46	-0.91	0.24	1.94	2.84	
$(\sqrt{d_i} - \overline{\sqrt{D}})^2$	2.92	3.84	0.21	0.82	0.058	3.77	8.06	$\sum = 19.678$
$(v_i - \overline{V})(\sqrt{d_i} - \overline{\sqrt{D}})$	40.3	40.31	6.24	4.16	1.54	31.87	111.98	$\sum = 236.4$

On a

- $V(V) = 435.66$ et $\sigma_V = 20.87$.
- $V(\sqrt{D}) = 2.81$ et $\sigma_{\sqrt{D}} = 1.67$.
- $cov(V, \sqrt{D}) = 33.77$.

Le coefficient de corrélation est
$$\rho(V, \sqrt{D}) = \frac{cov(V, \sqrt{D})}{\sigma_V \times \sigma_{\sqrt{D}}} = \frac{33.77}{20.87 \times 1.67} \simeq 0.968.$$
Droite de régression de $\sqrt{D}$ en V.
$$a = \frac{cov(V, \sqrt{D})}{V(V)} = \frac{33.77}{435.66} \simeq 0.077 \quad \text{et;} \quad b = \overline{\sqrt{D}} - a\overline{V} = 4.26 - 0.077 \times 53.57 = 0.135.$$
Ainsi, l'équation la droite de régression est $\boxed{(D_{\sqrt{D}/V}) : \quad \sqrt{D} = 0.077 \times V + 0.135.}$

5. Une voiture roulant à 100 km/h s'arrêtera au bout de $D = (0.077 \times 100 + 0.135)^2 = 61.38$ mètres.
6. Nous remarquons que $\rho(V, \sqrt{D})$ est plus proche de 1 que $\rho(V, D)$. Ainsi, la deuxième approximation semble être meilleure.
7. Une voiture qui s'arrête en 40 mètres aurait pour vitesse :
   - pour la $1^{\text{ère}}$ méthode : $V = \frac{40+18.16}{0.73} \simeq 79.67$.
   - pour la $2^{\text{ème}}$ méthode : $V = \frac{\sqrt{40}-0.135}{0.077} \simeq 80.38$.