



INSTITUT NATIONAL  
DES SCIENCES  
APPLIQUÉES  
TOULOUSE

Département STPI  
3ème année pré-orientation MIC  
Option GMM

## Statistique descriptive

Année universitaire 2020-2021

**Mélisande ALBERT**

Bureau 115 GMM  
[melisande.albert@insa-toulouse.fr](mailto:melisande.albert@insa-toulouse.fr)

Ce polycopié s'inspire des cours de Cathy Maugis-Rabusseau, Clément Marteau, Alain Baccini, Antoine Ayache et Julien Harmonier, et de Christine Tuleau-Malot. Les notions abordées sont également détaillées sur le Wikistat. Dans un but d'amélioration, je suis à l'écoute de tout commentaire constructif autant sur la forme que sur le fond. Merci d'avance.

### À propos de ce cours.

Deux parties : {

- (I) STATISTIQUE DESCRIPTIVE (évaluation par projets, en trinômes).
- (II) STATISTIQUE INFÉRENTIELLE (évaluation par CC).

Installer pour les TP : {

- LaTeX / MikTeX
- R
- RStudio

# Table des matières

<b>1. Introduction</b>	<b>5</b>
1.1. Vocabulaire . . . . .	5
1.2. Types de variables . . . . .	7
1.3. Un exemple illustratif : étude descriptive des pourboires . . . . .	8
<b>2. Statistique descriptive unidimensionnelle</b>	<b>9</b>
2.1. Variable qualitative . . . . .	9
2.1.1. Présentation . . . . .	9
2.1.2. Représentation graphique . . . . .	10
2.1.3. Indicateurs statistiques . . . . .	11
2.2. Variable quantitative . . . . .	11
2.2.1. Présentation . . . . .	11
2.2.2. Représentation graphique . . . . .	12
2.2.3. Indicateurs statistiques . . . . .	13
2.3. Exemples avec R . . . . .	18
<b>3. Statistique descriptive bidimensionnelle</b>	<b>23</b>
3.1. Deux variables quantitatives . . . . .	23
3.1.1. Représentation graphique . . . . .	23
3.1.2. Indices de liaison . . . . .	23
3.2. Une variable quantitative et une variable qualitative . . . . .	25
3.2.1. Représentation graphique . . . . .	25
3.2.2. Indice de liaison . . . . .	26
3.3. Deux variables qualitatives . . . . .	27
3.3.1. Représentation graphique . . . . .	28
3.3.2. Indices de liaison . . . . .	28
3.4. Exemples avec R . . . . .	29
<b>4. Analyse en Composantes Principales</b>	<b>31</b>
4.1. Introduction . . . . .	31
4.1.1. Positionnement du problème . . . . .	31
4.1.2. Transformation des données . . . . .	32
4.2. Inertie . . . . .	34
4.2.1. Représentation des individus et des variables . . . . .	34
4.2.2. Inertie globale . . . . .	35
4.2.3. Inertie axiale . . . . .	36
4.3. Recherche des composantes principales . . . . .	38
4.3.1. Première composante principale . . . . .	38

4.3.2. Décomposition en composantes principales . . . . .	39
4.4. Interprétation des résultats . . . . .	42
4.4.1. Choix du nombre de composantes . . . . .	42
4.4.2. Liens entre les individus et les axes principaux . . . . .	44
4.4.3. Liens entre les variables et les composantes principales . . . . .	45

# Chapitre 1

## Introduction

On appelle *Statistique* l'ensemble des méthodes permettant l'analyse des observations. Il est possible de classer les méthodes statistiques en deux groupes.

D'une part, l'objectif de la *statistique descriptive* est de décrire de façon claire et synthétique des données observées afin de mieux les analyser. Cette description se fait grâce à leur **présentation** (la plus commode), le calcul de **résumés numériques** et leur **représentation graphique**. Cette étude se fait sans aucune hypothèse de type probabiliste (par exemple, il n'est pas nécessaire de supposer que les données suivent une loi particulière).

D'autre part, la *statistique inférentielle* regroupe les méthodes dont l'objectif est d'induire les caractéristiques inconnues d'une population à partir d'observations, avec des marges d'erreurs contrôlées grâce à des hypothèses probabilistes. Généralement, la statistique descriptive précède la statistique inférentielle dans une démarche de traitement des données ; ces deux aspects de la statistique étant complémentaires.

**Exemple des mentions au baccalauréat.** Toutes les notions introduites dans ce chapitre seront illustrées sur un jeu de données (c.f. Table 1) contenant la couleur des yeux, le sexe, la taille, le poids, le nombre d'enfants et la mention au baccalauréat de 20 personnes. L'ensemble des analyses du cours sont faites avec le logiciel **R**.

**Objectifs du cours.** Les objectifs de ce cours sont multiples :

- ✓ Maîtriser les principales techniques de statistique descriptive (uni- et bi-dimensionnelles).
- ✓ Maîtriser les principes de l'Analyse en Composantes Principales (ACP).
- ✓ Être capable de mettre en œuvre ces techniques de manière adaptée au contexte de l'étude.
- ✓ Être capable d'appliquer ces techniques au moyen du logiciel **R** et savoir interpréter les sorties.

### 1.1 Vocabulaire

Avant d'aller plus loin, commençons par définir des notions que nous allons rencontrer dans toute la suite de ce cours.

**Population** : généralement notée  $\Omega$ , la *population* désigne l'ensemble concerné par l'étude statistique (ensemble d'éléments homogènes auxquels on s'intéresse). Dans l'exemple des mentions au baccalauréat, la population est constituée de l'ensemble des personnes ayant passé le baccalauréat. Si l'on s'intéresse à l'utilisation des VélôToulouse, la population est alors constituée de l'ensemble des vélos mis à disposition par le service. La notion de population est donc plus générale en statistique que dans le langage courant.

Prénom	Couleur des yeux	Sexe	Taille	Poids	Nb enfants	Mention
Guillaume	M	H	1.86	66.5	0	TB
Agnes	M	F	1.62	50.5	1	B
Thomas	M	H	1.72	67.5	2	B
Julie	B	F	1.67	52	2	AB
Sebastien	V	H	1.98	83	0	P
Stephanie	B	F	1.77	65	3	AB
Gregory	M	H	1.83	79	2	B
Anna	M	F	1.68	64	2	TB
Baptiste	B	H	1.92	81	3	B
Camille	V	F	1.71	53	1	AB
Frédéric	M	H	1.89	70	1	B
Emma	M	F	1.65	55	1	TB
Jules	M	H	1.75	68	2	B
Lucie	B	F	1.7	55	2	AB
Aymeric	B	H	2.1	95	2	P
Charlotte	B	F	1.77	68	3	P
Thibault	V	H	1.82	79	1	AB
Claire	V	F	1.5	50	3	P
Simon	B	H	1.95	90	3	B
Lise	M	F	1.74	60	0	TB

TABLE 1.1 – Caractéristique de 20 personnes ayant passé le baccalauréat.

**Individu (ou unité statistique)** : les éléments de la population sont appelés *individus*. Dans notre exemple, chaque individu correspond à un étudiant ayant passé le bac. Pour l'étude des Vélô-Toulouse, chaque individu correspond à un vélo.

**Échantillon** : c'est la partie de la population qui participe effectivement à l'étude menée (sur laquelle sont réalisées les observations). En effet, il est fréquent que l'on ne puisse pas observer la population toute entière. Dans notre exemple, il s'agit des 20 personnes interrogées. Son cardinal est appelé *taille de l'échantillon*.

**Variable statistique** : cela correspond à une caractéristique définie sur la population, observée sur chaque individu de l'échantillon. Par exemple, le jeu de données étudié ici comprend 6 variables qui sont la couleur des yeux, le sexe, la taille, le poids, le nombre d'enfants et la mention au bac.

**Modalité (levels en anglais)** : les *modalités* d'une variable statistique sont les valeurs possibles de cette variable. Par exemple, la variable "couleur des yeux" prend trois modalités, à savoir "M" pour marron, "V" pour vert et "B" pour bleu. Pour les variables "Taille" et "Poids" (respectivement "Nb.enfants"), les modalités sont tous les réels positifs (respectivement les entiers naturels). Finalement, la variable "Mention" prend 4 modalités qui sont "TB" si l'individu a eu la mention "Très bien", "B" pour "Bien", "AB" pour "Assez bien" et "P" pour "Passable".

On remarquera que chaque individu appartient à une seule modalité. En effet, on ne peut avoir des individus ayant les yeux de plusieurs couleurs, ou plusieurs mentions au baccalauréat.

**Données** : cela désigne l'ensemble des observations des différentes variables considérées sur chaque individu de l'échantillon. Les données sont généralement présentées sous forme de tableau avec les individus en lignes, et les variables en colonnes.

**Série statistique** : On appelle *série statistique* la suite des valeurs prises par la variable étudiée. Cela correspond à la suite des valeurs observées sur l'échantillon.

## 1.2 Types de variables

L'étude statistique d'un jeu de données dépend fortement de la nature des variables étudiées. On distingue 4 types de variables statistiques :

- Une variable est dite ***quantitative*** si ses valeurs possibles sont des valeurs numériques. Par exemple, pour les VéloToulouse, les variables "hauteur de la selle en fin de journée" ou "nombre d'utilisations" sont des variables quantitatives. Ces variables peuvent être classées en deux types :
  - Une variable est dite **quantitative discrète** si l'ensemble de ses modalités est au plus dénombrable (par exemple le nombre d'utilisations).
  - Une variable est dite **quantitative continue** si l'ensemble de ses modalités est continu (par exemple la hauteur de la selle en fin de journée).
- Une variable est dite ***qualitative*** si ses modalités sont désignées par des noms (ou des catégories). Par exemple, l'année de mise en service, ou le nom de la station en fin de journée.
  - Si ses modalités peuvent être classés, alors la variable est dite **qualitative ordinaire** (par exemple, l'année de mise en service).
  - Sinon, la variable est dite **qualitative nominale** (par exemple le nom de la station en fin de journée).

Remarquons que pour les variables quantitatives, il est possible d'additionner les observations (et en particulier, de calculer des moyennes, etc.). Par exemple, la variable "poids" est quantitative puisque, si plusieurs individus montent sur une balance, leur masse totale est la somme de leurs masses individuelles.

*A contrario*, même si la variable "numéro de téléphone" d'un ensemble de personnes est identifiée à un nombre, elle ne représente pas une quantité : on ne peut pas faire un total avec les numéros de téléphone de plusieurs individus (du moins, le résultat d'un tel calcul n'aurait aucun sens).

**Exemple :** Donner la nature (ou le type) des variables statistiques suivantes en précisant, dans le cas de variables qualitatives, le nombre de modalités :

- la couleur des yeux : .....
- le sexe : .....
- la taille : .....
- le nombre d'enfants : .....
- la mention au bac : .....

Voici le résultat obtenu avec le logiciel  :

```
> Data<-read.table("DonneesCoursSD.txt",header=T)

> dim(Data)
[1] 20  6

> str(Data)
'data.frame': 20 obs. of  6 variables:
 $ Couleur.des.yeux: Factor w/ 3 levels "B","M","V": 2 2 2 1 3 1 2 2 1 3 ...
 $ Sexe            : Factor w/ 2 levels "F","H": 2 1 2 1 2 1 2 1 2 1 ...
 $ Taille           : num  1.86 1.62 1.72 1.67 1.98 1.77 1.83 1.68 1.92 1.71 ...
 $ Poids            : num  66.5 50.5 67.5 52 83 65 79 64 81 53 ...
 $ Nb.enfants       : int  0 1 2 2 0 3 2 2 3 1 ...
 $ Mention          : Factor w/ 4 levels "AB","B","P","TB": 4 2 2 1 3 1 2 4 2 1 ...
```

D'une part, nous pouvons remarquer que les variables qualitatives sont codées au format `Factor`. D'autre part, les variables quantitatives sont codées au format `int` (pour *integer* en anglais) si elles sont discrètes, et `num` (pour *numeric*) si elles sont continues. Avant de commencer une étude statistique avec , il faut donc toujours s'assurer que le format des variables est le bon.

### 1.3 Un exemple illustratif : étude descriptive des pourboires

Ce tableau contient les pourboires (*tips* en anglais) d'un serveur dans un restaurant américain aux débuts des années 1990.

Le restaurant était dans un centre commercial. Les données indiquent respectivement le prix du repas, le pourboire, le genre de la personne qui a payé et donné le pourboire, le nombre de convives et enfin le degré de satisfaction (I : insatisfait, S : satisfait, TS : très satisfait) .

ID	TOTBILL	TIP	SEX	SIZE	SATISF
185	40.55	3.00	H	2	S
184	23.17	6.50	H	4	TS
1	16.99	1.01	F	2	I
166	24.52	3.48	H	3	S
163	16.21	2.00	F	3	I
160	16.49	2.00	H	4	TS
179	9.60	4.00	F	2	TS
151	14.07	2.50	H	2	S
117	29.93	5.07	H	4	S
19	16.97	3.50	F	3	S

Nous allons mener à la main, l'étude descriptive uni- et bi-dimensionnelle de ce jeu de données afin d'illustrer les différentes notions rencontrées dans le cours.

# Chapitre 2

## Statistique descriptive unidimensionnelle

La première étape en statistique descriptive consiste à étudier séparément chaque variable du jeu de données : il s'agit de l'*étude unidimensionnelle*. Comme cela a été précisé précédemment, les méthodes utilisées pour décrire correctement un jeu de données diffèrent selon le type de la variable ; il est donc nécessaire d'avoir bien identifié la nature de la variable étudiée au préalable.

Dans cette section, nous nous intéressons à l'étude d'une variable statistique, notée  $X$ . Notons  $n$  la taille de l'échantillon,  $x_1, \dots, x_n$  les valeurs observées de la variable  $X$ , et  $\underline{x} = (x_1, \dots, x_n)$  la série statistique. Les différentes techniques présentées dans cette section (présentation, représentation graphique, calcul de caractéristiques numériques) sont classées selon le type de la variable  $X$ .

### 2.1 Variable qualitative

Supposons que la variable étudiée  $X$  est qualitative, admettant  $K$  modalités, notées  $m_1, \dots, m_K$ .

#### 2.1.1 Présentation

##### a) Cas d'une variable qualitative nominale

Supposons que  $X$  est une variable qualitative *nominale*, c'est-à-dire que ses modalités  $m_1, \dots, m_K$  ne peuvent pas être classées dans un ordre naturel. L'observation de cette variable  $X$  sur les  $n$  individus (de l'échantillon) peut se résumer par un tableau des effectifs  $n_k$ , ou des fréquences  $f_k$  définis par

$$n_k = \sum_{i=1}^n \mathbb{1}_{\{x_i=m_k\}} \quad \text{et} \quad f_k = \frac{n_k}{n}.$$

Remarquons que  $\sum_{k=1}^K n_k = n$ , et que, par conséquent,  $\sum_{k=1}^K f_k = 1$ .

Modalité	Effectif	Fréquence
$m_1$	$n_1$	$f_1$
$m_2$	$n_2$	$f_2$
$\vdots$	$\vdots$	$\vdots$
$m_K$	$n_K$	$f_K$

##### b) Cas d'une variable qualitative ordinale

Supposons maintenant que  $X$  est une variable qualitative *ordinale*. Ses modalités peuvent donc être ordonnées

$$m_1 \prec m_2 \prec \dots \prec m_K.$$

Comme précédemment, on peut s'intéresser aux effectifs  $n_k$  et fréquences  $f_k$ . On peut également considérer les effectifs cumulés  $N_k$  et les fréquences cumulées  $F_k$ , définis par

$$N_k = \sum_{j=1}^k n_j \quad \text{et} \quad F_k = \sum_{j=1}^k f_j.$$

Notons que  $N_K = n$  et  $F_K = 1$ .

Modalité	Eff.	Eff. cum.	Fréq.	Fréq. cum.
$m_1$	$n_1$	$N_1$	$f_1$	$F_1$
$m_2$	$n_2$	$N_2$	$f_2$	$F_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m_K$	$n_K$	$N_K$	$f_K$	$F_K$

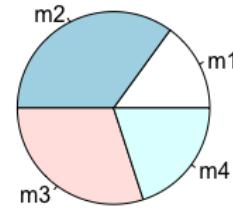
## 2.1.2 Représentation graphique

Les représentations graphiques des variables qualitatives sont nombreuses, mais ont le même principe général : les différentes modalités de la variable sont représentées par des parties du graphique dont la surface est proportionnelle à l'effectif (ou à la fréquence).

### a) Cas d'une variable qualitative nominale

Une représentation graphique adaptée pour une **variable qualitative nominale** est le *diagramme en secteurs*, ou *diagramme circulaire (pie chart)* en anglais, ou encore *camembert* en français). Elle permet de faire apparaître la répartition des fréquences entre les différentes modalités. Chaque angle  $\alpha_k$ , correspondant à la modalité  $m_k$  est calculé en degrés par

$$\alpha_k = f_k \times 360.$$



### b) Cas d'une variable qualitative ordinaire

Pour une **variable qualitative ordinaire**, les représentations graphiques les plus adaptées sont

- les *diagramme en bâtons* des fréquences (cumulées ou non) (*bar chart* en anglais),
- ou le *diagramme en barre* (*line chart* en anglais).

Remarquons que, même si les diagrammes en bâtons des effectifs sont également possibles, ceux des fréquences sont privilégiés car ils permettent notamment la comparaison aisée d'échantillons de taille différentes. Par ailleurs, il est conventionnel que sur l'axe des abscisses, les modalités apparaissent selon l'ordre implicite. Cependant, il faut bien faire attention à ne pas graduer l'axe des abscisses car l'ordre n'est qu'implicite et l'écart entre deux modalités n'est pas significatif.

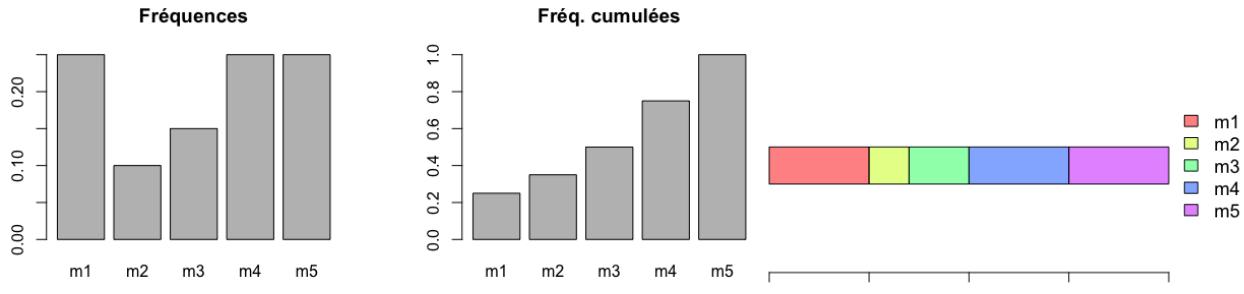


Diagramme en *bâtons* des fréquences (gauche) et des fréquences cumulées (droite) d'une variable qualitative ordinaire

Diagramme en *barre* d'une variable qualitative ordinaire

Notons finalement que, même si dans la théorie, chacun des graphiques présentés ci-dessus est bien défini quel que soit le type de variable qualitative (nominal ou ordinal), nous n'échangerons pas les représentations, car on perd la relation d'ordre dans le diagramme en secteurs alors que le diagramme en bâtons en crée une.

### 2.1.3 Indicateurs statistiques

Il n'existe qu'un seul indicateur statistique pour les variables qualitatives : le *mode* est la valeur la plus fréquente de l'échantillon.

Notons que le mode n'est pas nécessairement unique ;

- en cas d'unicité on parle de distribution unimodale,
- sinon, on parle de distribution plurimodale.

Il est clair qu'il n'est pas possible de calculer de caractéristiques numériques de position ou de dispersion pour des variables qualitatives.

## 2.2 Variable quantitative

### 2.2.1 Présentation

#### a) Cas d'une variable quantitative discrète

Supposons à présent que  $X$  est une variable quantitative *discrète*. Ses modalités (au plus dénombrables) sont des nombres réels et sont donc ordonnées

$$m_1 \leq m_2 \leq \dots \leq m_k \leq \dots$$

Comme pour les variables qualitatives ordinaires, on peut s'intéresser aux effectifs  $n_k$ , aux effectifs cumulés  $N_k$ , aux fréquences  $f_k$  et aux fréquences cumulées  $F_k$  des modalités (jusqu'à la  $K$ ème plus grande modalité observée), et les représenter dans un tableau.

Modalité	Effectifs	Eff. cumulés	Fréquences	Fréq. cumulées
$m_1$	$n_1$	$N_1$	$f_1$	$F_1$
$m_2$	$n_2$	$N_2$	$f_2$	$F_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m_K$	$n_K$	$N_K$	$f_K$	$F_K$

#### b) Cas d'une variable quantitative continue

Dans le cas où  $X$  est de nature quantitative continue, sa représentation est plus compliquée. En effet, il est impossible d'énumérer l'ensemble des valeurs possibles, car dans la théorie, il s'agit de tous les réels à l'intérieur d'un intervalle. En particulier, cela rend impossible de lister les différentes modalités dans la première colonne comme dans les autres cas. Il faut donc constituer des classes  $[a_k, a_{k+1}[$  qui serviront de références pour le tableau de représentation. On peut donc calculer les effectifs (cumulés ou non) de chaque classe et les fréquences (cumulées ou non) par classe :

$$n_k = \sum_{i=1}^n \mathbb{1}_{\{a_k \leq x_i < a_{k+1}\}}, \quad N_k = \sum_{j=1}^k n_j, \quad f_k = \frac{n_k}{n} \quad \text{et} \quad F_k = \sum_{j=1}^k f_j,$$

et les représenter dans un tableau.

Classe	Effectifs	Eff. cumulés	Fréquences	Fréq. cumulées
$[a_1, a_2[$	$n_1$	$N_1$	$f_1$	$F_1$
$[a_2, a_3[$	$n_2$	$N_2$	$f_2$	$F_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_K, a_{K+1}[$	$n_K$	$N_K$	$f_K$	$F_K$

Remarque : Si l'on étudie une variable quantitative discrète pouvant prendre un grand nombre de valeurs, il est parfois préférable de faire comme pour les variables continues, et de regrouper les valeurs en classes.

Il se peut que, pour des raisons pratiques, ou si l'on étudie une variable à caractère sensible, les données sont déjà données sous forme d'intervalles. Par exemple, dans une enquête, il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis ; de même pour l'âge.

Cependant, il arrive fréquemment que ce ne soit pas le cas. Alors, une question fondamentale se pose : *comment construire des classes qui soient adaptées au jeu de données ?* Il existe plusieurs stratégies, parfois basées sur des mesures de dispersion (définies dans la partie 2.2.3) dont trois sont données ci-dessous pour information, et sont implémentées dans **R**.

- La règle de **Sturges** détermine empiriquement le nombre de classes :

$$K = \lceil 1 + \log_2(n) \rceil,$$

où  $n$  représente la taille de l'échantillon, et  $\lceil \cdot \rceil$  la valeur entière supérieure (pour que  $K$  soit un entier).

- La règle de **Scott** dépend de l'écart-type (empirique)  $s_x$  des données et fixe la longueur des classes à  $a_{k+1} - a_k = 3.5 \times s_x / \sqrt[3]{n}$ .
- La règle de **Freedman-Diaconis** dépend de l'écart inter-quartiles  $IQ(x)$  fixe la longueur des classes à  $2 \times IQ(x) / \sqrt[3]{n}$ .

On rappelle que les quantités  $s_x$  et  $IQ(x)$  sont définies dans la partie 2.2.3.

## 2.2.2 Représentation graphique

### a) Cas d'une variable quantitative discrète

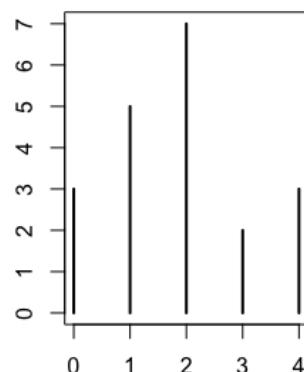
Dans ce cas, les représentations graphiques les plus adaptées sont

- le *diagramme en bâtons* des fréquences (ou des effectifs) : contrairement au cas des variables qualitatives ordinaires, cette fois-ci, les modalités sont ordonnées sur l'axe des abscisses.
- la *courbe des fréquences cumulées* : il s'agit du graphe de la *fonction de répartition empirique* (*empirical cumulative distribution function* en anglais) définie pour tout  $t$  dans  $\mathbb{R}$  par

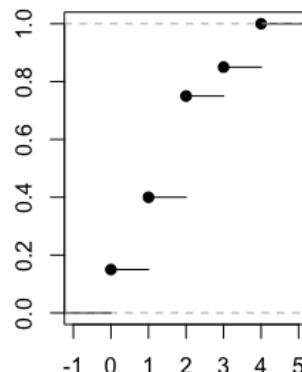
$$F_x(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{x_i \leq t\}} = \begin{cases} 0 & \text{si } t < m_1 \\ F_k & \text{si } m_k \leq t < m_{k+1} \\ 1 & \text{si } t \geq m_K. \end{cases}$$

La variable étant discrète, cette courbe est constante par morceaux, continue à droite et limitée à gauche (càdlàg). Par ailleurs, elle est croissante et vérifie  $\lim_{t \rightarrow -\infty} F_x(t) = 0$  et  $\lim_{t \rightarrow +\infty} F_x(t) = 1$ .

Diagramme bâton



Fct. répartition



### b) Cas d'une variable quantitative continue

Dans ce cas, les représentations graphiques les plus adaptées sont

- l'*histogramme* : il correspond au diagramme en bâtons dans les autres cas, sauf qu'ici, les bâtons sont bien évidemment collés puisqu'il existe une continuité entre les valeurs en abscisse.

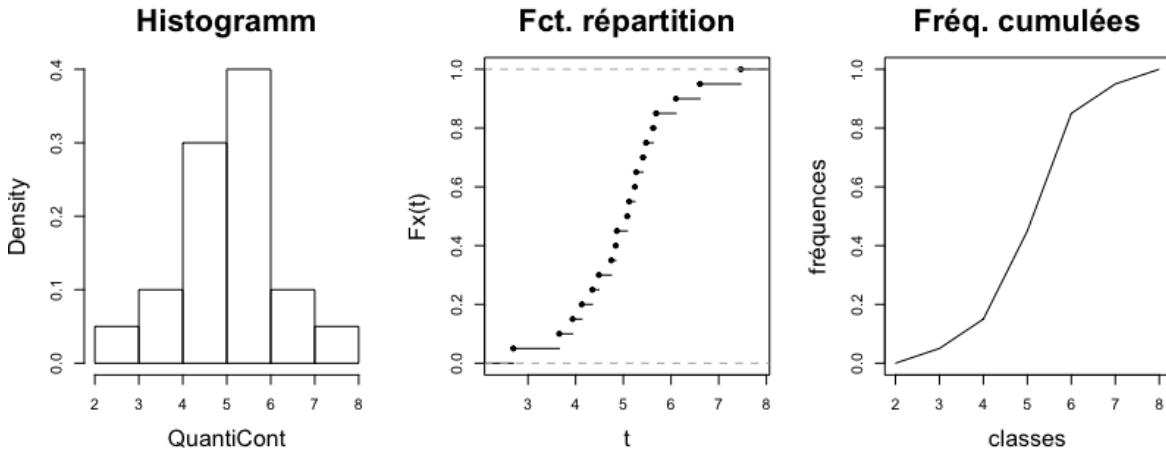
La hauteur associée à la  $j$ ème classe  $[a_k, a_{k+1}]$  est calculée par

$$h_k = \frac{f_k}{a_{k+1} - a_k}.$$

Attention à ne pas oublier la renormalisation par l'amplitude de la classe, indispensable dans le cas où les classes ne sont pas de même longueur ! En effet, lorsqu'on s'intéresse à l'aspect modélisation, on cherche, par exemple, une loi dont la densité se rapproche de l'allure de notre histogramme. Or, cette renormalisation permet d'obtenir un histogramme d'aire égale à 1 (comme pour les densités).

- la *courbe de la fonction de répartition empirique* : comme pour les variables discrètes, on peut tracer la fonction de répartition empirique à partir de chaque observation (sans tenir compte des classes).
- la *courbe des fréquences cumulées* : la fonction de répartition (théorique) d'une variable aléatoire continue étant continue, il peut être préférable de faire une approximation par une fonction linéaire par morceaux, en tenant compte des classes. Pour cela, on place sur le graphe les points de coordonnées  $(a_1, 0)$  et  $(a_{k+1}, F_k)$  pour  $k$  allant de 1 à  $K$ , puis on relie les points par des segments de droites.

Il faut cependant garder en mémoire que cette représentation (linéaire par morceaux) est basé sur une hypothèse très forte quant à la répartition des données (qui n'est pas faite avec la fonction de répartition). En effet, cela revient à supposer que les données sont réparties de façon homogène à l'intérieur d'une classe. On sait dans la réalité que ce n'est pas le cas, mais cela demeure une approximation usuelle.



### 2.2.3 Indicateurs statistiques

Dans toute la suite, sauf précision du contraire, les indicateurs statistiques sont empiriques, c'est-à-dire, calculés à partir les observations. Attention à ne pas les confondre avec les indicateurs statistiques théoriques, dépendant généralement d'hypothèses de loi sur les variables dont les caractéristiques ne sont pas connues.

#### a) Mesures de tendance centrale

**Le mode.** Comme dans le cas d'une variable qualitative, on définit le *mode* d'une variable quantitative discrète comme la modalité la plus fréquente dans l'échantillon. Dans le cas d'une variable

quantitative continue (ou discrète regroupée par classes), on définit la notion de *classe modale*. C'est la classe pour laquelle la hauteur dans l'histogramme (à savoir la fréquence divisée par l'amplitude de la classe) est la plus élevée.

Attention, il ne s'agit pas nécessairement de la classe associée à l'effectif le plus élevé ! En effet, cette seconde définition n'est vraie que pour des classes de même amplitude, mais devient complètement erronée dans le cadre de classes d'amplitudes distinctes.

Remarquons que le mode a l'avantage d'être simple à calculer, et de ne pas dépendre des valeurs extrêmes. Cependant, comme précisé précédemment, il peut y en avoir plusieurs. En outre, plus ils sont nombreux, moins ils ont d'importance. Par ailleurs, le mode ne donne pas de renseignement sur l'ensemble des données.

**La moyenne.** La *moyenne* (empirique) est définie par  $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ .

Notons qu'elle repose sur la somme des observations, et donc n'est évidemment pas définie pour des variables qualitatives. Elle peut être calculée via les modalités observées et leur effectif (ou leur fréquence) :

$$\bar{x}_n = \frac{1}{n} \sum_{k=1}^K n_k m_k = \sum_{k=1}^K f_k m_k.$$

En effet,

$$\begin{aligned} \bar{x}_n &= \frac{1}{n} \sum_{i=1}^n \left[ x_i \times \underbrace{\sum_{k=1}^K \mathbb{1}_{x_i = m_k}}_1 \right] \\ &= \sum_{k=1}^K \left[ \frac{1}{n} \sum_{i=1}^n \underbrace{x_i \mathbb{1}_{x_i = m_k}}_{m_k} \right] = \sum_{k=1}^K \left[ m_k \times \frac{1}{n} \underbrace{\sum_{i=1}^n \mathbb{1}_{x_i = m_k}}_{\text{nb d'individus qui ont pris la modalité } m_k} \right] \end{aligned}$$

Plus généralement, il est possible d'accorder des poids différents pour chacune des observations. Pour cela, considérons pour chaque observation  $i$  le poids correspondant  $0 \leq \omega_i \leq 1$ , en supposant que  $\sum_{i=1}^n \omega_i = 1$ . La **moyenne pondérée** est alors définie par

$$\bar{x}_\omega = \sum_{i=1}^n \omega_i x_i.$$

On retrouve bien la moyenne en prenant  $\omega_i = 1/n$  pour tout  $i$ .

La moyenne a l'avantage de tenir compte de l'ensemble des données. L'inconvénient est qu'elle est très dépendante des valeurs extrêmes. C'est généralement la tendance centrale la plus populaire des trois présentées ici (mode, moyenne, médiane), même si parfois, elle n'est pas représentative.

Finalement, on dira que l'on *centre la variable* si l'on soustrait la moyenne à chaque observation, c'est-à-dire que l'on étudie plutôt  $x_i - \bar{x}_n$ .

**La médiane.** Par définition, la *médiane* (empirique) est une valeur qui divise l'échantillon en deux sous-échantillons de même cardinal : au moins la moitié des valeurs observées est supérieure ou égale à la médiane, et au moins la moitié des valeurs observées est inférieur ou égale à la médiane :

$$\sum_{i=1}^n \mathbb{1}_{\{x_i \geq m\}} \geq \frac{n}{2} \quad \text{et} \quad \sum_{i=1}^n \mathbb{1}_{\{x_i \leq m\}} \geq \frac{n}{2}.$$

Le choix de la médiane est multiple (plus de détails sont donnés dans la partie sur les quantiles ci-dessous). On peut retenir que

- si  $X$  est discrète, alors,  $m$  est la première modalité associée à la fréquence cumulée qui est immédiatement supérieure ou égale à 0.5.
- si  $X$  est continue, il existe plusieurs choix, tenant compte ou non des classes, et si oui, prenant le bord ou le centre de la classe.

La médiane a le gros avantage de ne pas dépendre des valeurs extrêmes. En particulier, elle est plus représentative que la moyenne si le jeu de données présente des valeurs extrêmes et que la distribution est asymétrique. Cependant, tout comme le mode, elle ne donne pas de renseignement sur l'ensemble des données.

Remarquons finalement qu'un grand écart entre la moyenne et la médiane peut indiquer la possibilité de présence de données extrêmes. La réciproque est fausse.

### b) Mesures de position

**Les quartiles.** Un *quartile* est chacune des trois valeurs qui divisent les données triées en quatre parts égales, de sorte que chaque partie contient 25% de l'échantillon :

- Le premier quartile, noté  $q_{0.25}$ , est une valeur qui sépare les 25% des valeurs inférieures de l'échantillon :
  - au moins 25% de l'échantillon prend des valeurs inférieures ou égales à  $q_{0.25}$ ,
  - et au moins 75% de l'échantillon prend des valeurs supérieures ou égales.
- Le deuxième quartile  $q_{0.5}$  est une valeur qui sépare l'échantillon en deux parties de même taille, (c'est la médiane).
- Le troisième quartile  $q_{0.75}$  est une valeur qui sépare les 25% des valeurs supérieures de l'échantillon.

**Les quantiles.** Ils généralisent les notions de médiane/quartiles à n'importe quelle proportion  $\alpha$  dans  $]0, 1[$ . Rappelons que si  $X$  est une variable aléatoire de fonction de répartition  $F_X$  (vrai, mais en général inconnue), le  $\alpha$ -quantile théorique est défini par

$$Q_\alpha = \inf\{t \in \mathbb{R}; F_X(t) \geq \alpha\}.$$

Il est donc naturel de définir le  $\alpha$ -quantile empirique à partir de la fonction de répartition empirique  $F_{\underline{x}}$  par

$$\begin{aligned} q_\alpha &= \inf\{t \in \mathbb{R}; F_{\underline{x}}(t) \geq \alpha\} \\ &= x_{(i)} \quad \text{avec } \alpha \in \left[\frac{i-1}{n}; \frac{i}{n}\right] \end{aligned}$$

où  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  sont les valeurs ordonnées de la série statistique.

Remarquons que cette définition peut être critiquable.

Par exemple, considérons  $\underline{x} = (1, 2, 3, 7, 8, 9)$  et  $\alpha = \frac{1}{2}$ .

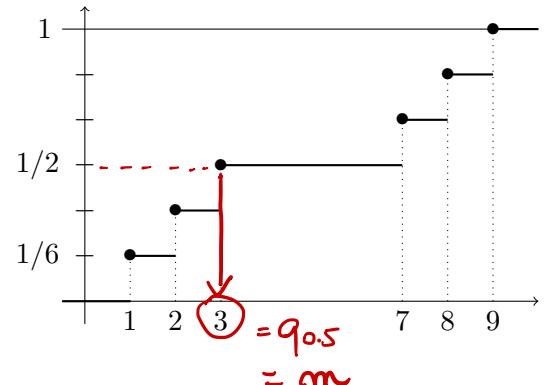
D'après la définition,  $q_{0.5} = 3$  mais

$$\forall t \in [3, 7[, \quad F_{\underline{x}}(t) = \frac{1}{2}.$$

On peut préférer comme définition pour  $q_\alpha$  le centre de l'intervalle, ici  $q_{0.5} = 5$ .

Il existe plusieurs cas particuliers :

- la médiane correspond au quantile d'ordre  $\alpha = 0.5$ .
- le premier et le troisième quartiles correspondent à  $\alpha = 0.25$  et  $\alpha = 0.75$  respectivement.
- de même, on peut définir les *déciles* ou les *centiles* en prenant pour  $\alpha$  des multiples de 0.1 ou de 0.01.



### Représentation graphique : la boîte à moustaches (*boxplot* en anglais).

La boîte à moustaches est un graphique qui résume la série statistique à partir de ses valeurs extrêmes, ses quartiles. Plus précisément, elle représente :

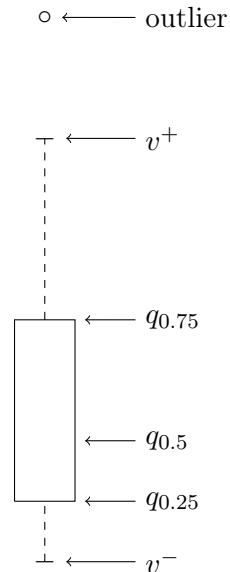
- les quartiles (dont la médiane),
- la *valeur adjacente supérieure*  $v^+$  est la plus grande valeur de l'échantillon inférieure ou égale à

$$L_+ := q_{0.75} + 1.5 \times (q_{0.75} - q_{0.25}),$$

- la *valeur adjacente inférieure*  $v^-$  est la plus petite valeur de l'échantillon supérieure ou égale à

$$L_- := q_{0.25} - 1.5 \times (q_{0.75} - q_{0.25}),$$

- les valeurs extrêmes (*outliers* en anglais) sont les valeurs de l'échantillon n'appartenant pas à  $[v^-, v^+]$ .



### c) Mesures de dispersion

**Étendue.** L'*étendue* est la différence entre la plus grande et la plus petite des valeurs observées

$$x_{(n)} - x_{(1)}.$$

Remarquons que l'*étendue* a l'avantage d'être très simple à calculer. Cependant, elle dépend fortement des valeurs extrêmes et ne donne aucune information sur la dispersion des valeurs intermédiaires.

**Variance et écart-type.** La *variance* (empirique) est la moyenne (empirique) des carrés des écarts à la moyenne (empirique).

$$s_x^2 := \text{Var}(\underline{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x}_n)^2.$$

Remarquons que la variance est toujours positive (ou nulle) en tant que moyenne de carrés (toujours positifs ou nuls). On peut donc définir l'*écart-type* comme la racine de la variance :  $s_x = \sqrt{s_x^2}$ . Finalement, notons que la variance (et donc l'*écart-type*) est nulle si et seulement si tous les carrés de la somme sont nuls, c'est-à-dire que tous les  $x_i$  sont égaux. Si ce n'est pas le cas, on dit que l'on *centre et réduit la variable* si l'on soustrait la moyenne à chaque observation et on divise par l'*écart-type*, c'est-à-dire que l'on étudie plutôt

$$\tilde{x}_i = \frac{x_i - \bar{x}_n}{s_x}.$$

La variance et l'*écart-type* ont l'avantage de prendre en compte toutes les données. Ce sont des mesures absolues dans le sens où elles possèdent une unité (unité des données au carré pour la variance, et unité des données pour l'*écart-type*).

On notera que, malgré les apparences, la formule de variance est très naturelle. En effet, un indicateur de dispersion doit résumer l'*écartement* des observations autour de leur tendance centrale. Ici, il s'agit de mesures de dispersion vis à vis de la moyenne. Il est possible de prendre en compte d'autres mesures centrales menant à d'autres écarts représentatifs, tels que

- l'*écart moyen absolu* (moyenne des écarts absolus à la moyenne) :  $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}_n|$ ,
- l'*écart médian absolu* (moyenne des écarts absolus à la médiane) :  $\frac{1}{n} \sum_{i=1}^n |x_i - q_{0.5}|$ ,

**Écart inter-quantiles.** l'*écart inter-quartiles* est la différence des troisième et premier quartiles :

$$IQ(x) = q_{0.75} - q_{0.25}.$$

Il joue un peu le rôle de l'écart-type lorsque la médiane est plus représentative que la moyenne.

#### d) Autres mesures

**Moments.** Généralisant la variance, on peut également définir les moments (à n'importe quel ordre) de la manière suivante.

- Moment à l'origine d'ordre  $r \in \mathbb{N}$  :  $M'_r = \frac{1}{n} \sum_{i=1}^n x_i^r$
- Moment centré d'ordre  $r \in \mathbb{N}$  :  $M_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^r$ .

Notons que pour  $r = 1$  ou  $2$ , on retrouve la moyenne et la variance :

$$M'_1 = \bar{x}_n, \quad M_1 = 0, \quad M'_2 = s_x^2 + \bar{x}_n^2, \quad M_2 = s_x^2.$$

#### Cas particuliers : paramètres de forme.

- Le premier paramètre de forme est le *coefficient d'asymétrie de Pearson* (*skewness* en anglais). Il est défini par le moment d'ordre 3 de la variable centrée réduite :

$$\gamma_1 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}_n}{s_x} \right)^3 = \frac{M_3}{M_2^{3/2}}.$$

Un coefficient positif indique une distribution décalée à gauche de la médiane, et donc une queue de distribution étalée vers la droite, alors qu'en inversement, un coefficient négatif indique une queue de distribution étalée vers la gauche.

- Le second paramètre de forme est le *coefficient d'aplatissement de Pearson* (*kurtosis* en anglais). Il est défini par le moment d'ordre 4 de la variable centrée réduite :

$$\beta_n = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}_n}{s_x} \right)^4 = \frac{M_4}{M_2^2}.$$

On définit alors l'excès d'aplatissement (*excess kurtosis* en anglais) par

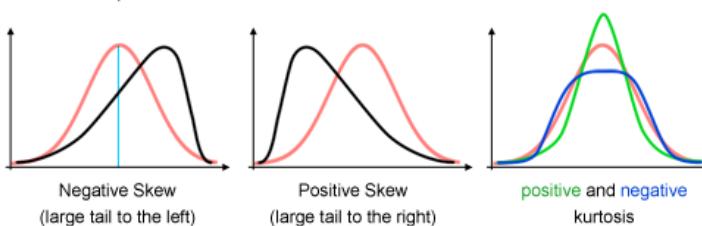
$$\gamma_2 = \beta_2 - 3$$

Remarquons que le "3" correspond au kurtosis (non normalisé) théorique d'une  $\mathcal{N}(0, \sigma^2)$ . Le coefficient d'excès d'aplatissement mesure donc l'excès d'aplatissement de la variable comparé à une loi Gaussienne.

Un coefficient d'aplatissement élevé indique que la distribution est plutôt pointue en sa moyenne, et a des queues de distribution épaisses.

Attention, le terme d'"excès d'aplatissement" (dérivé de *kurtosis excess* en anglais), utilisé pour le kurtosis normalisé peut être source d'ambiguïté. En effet, un excès d'aplatissement positif correspond à une distribution pointue et un excès d'aplatissement négatif à une distribution aplatie (on s'attendrait à l'inverse).

© www.scratchapixel.com



## 2.3 Exemples avec R

### VARIABLE COULEUR DES YEUX (qualitative nominale)

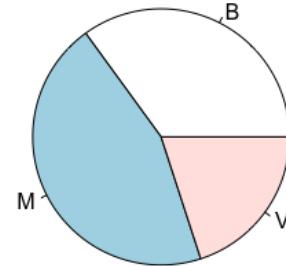
```
> Yeux<-Data[,1]
> class(Yeux)
[1] "factor"

> levels(Yeux)
[1] "B" "M" "V"

> V<-table(Yeux)

> data.frame(Eff=c(V), Freq=c(V)/sum(V))
  Eff Freq
B    7 0.35
M    9 0.45
V    4 0.20

> pie(table(Yeux))
```



### VARIABLE MENTION AU BAC (qualitative ordinale)

```
> Bac<-Data$Mention

> BacF=factor(Bac,levels=c("P","AB","B","TB"))

> B1=table(BacF)

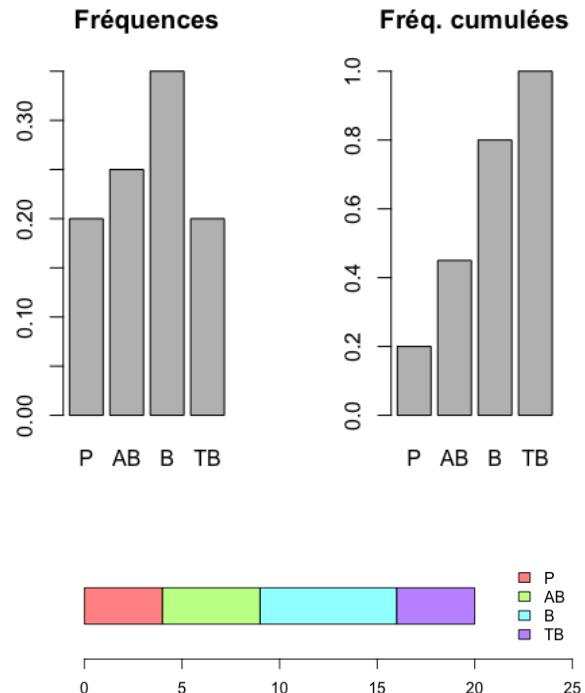
> data.frame(Eff=c(B1),
+             EffCum=cumsum(B1),
+             Freq=c(B1)/sum(B1),
+             FreqCum=cumsum(B1)/sum(B1))
  Eff EffCum Freq FreqCum
P     4      4 0.20   0.20
AB    5      9 0.25   0.45
B     7     16 0.35   0.80
TB    4     20 0.20   1.00

> par(mfrow=c(1,2))

> barplot(table(BacF)/sum(table(BacF)),
+         main="Fréquences")

> barplot(cumsum(table(BacF))/sum(table(BacF)),
+         main="Fréq. cumulées")

> barplot(as.matrix(table(BacF)),horiz=T,asp=2,
+          col=rainbow(4,alpha=.5),xlim=c(0,25))
> legend(x="right", legend=levels(BacF), cex=1,
+          fill=rainbow(4,alpha=.5),byt="n")
```



## VARIABLE "NB ENFANTS"

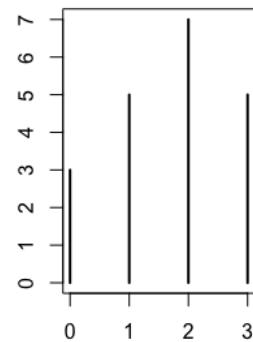
```
> NbEnfant<-Data[,"Nb.enfants"]

> Eff<-c(table(sort(NbEnfant)))

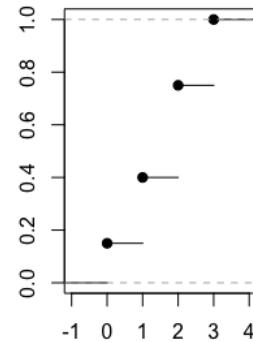
> data.frame(Eff=Eff,
+             EffCum=cumsum(Eff),
+             Freq=Eff/sum(Eff),
+             FreqCum=cumsum(Eff)/sum(Eff))

  Eff EffCum Freq FreqCum
0    3      3 0.15   0.15
1    5      8 0.25   0.40
2    7     15 0.35   0.75
3    5     20 0.25   1.00
```

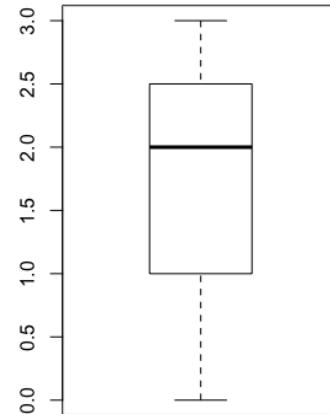
Diagramme bâton



Fct. répartition



Boxplot



```
> summary(NbEnfant)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
  0.00    1.00    2.00  1.70    2.25   3.00

> B.NbEnfant = boxplot(NbEnfant,main="Boxplot")

> B.NbEnfant$stats
  [,1]
[1,]  0.0
[2,]  1.0
[3,]  2.0
[4,]  2.5
[5,]  3.0
```

**VARIABLE "TAILLE"**

```
> Taille<-Data[, "Taille"]

> A<-hist(Taille,main="Histogramme des Eff.")

> A$breaks # par défaut Sturges, "Scott", "FD
[1] 1.5 1.6 1.7 1.8 1.9 2.0 2.1

> A$counts
[1] 1 5 6 4 3 1

> A$density
[1] 0.5 2.5 3.0 2.0 1.5 0.5

> A$mids
[1] 1.55 1.65 1.75 1.85 1.95 2.05
```

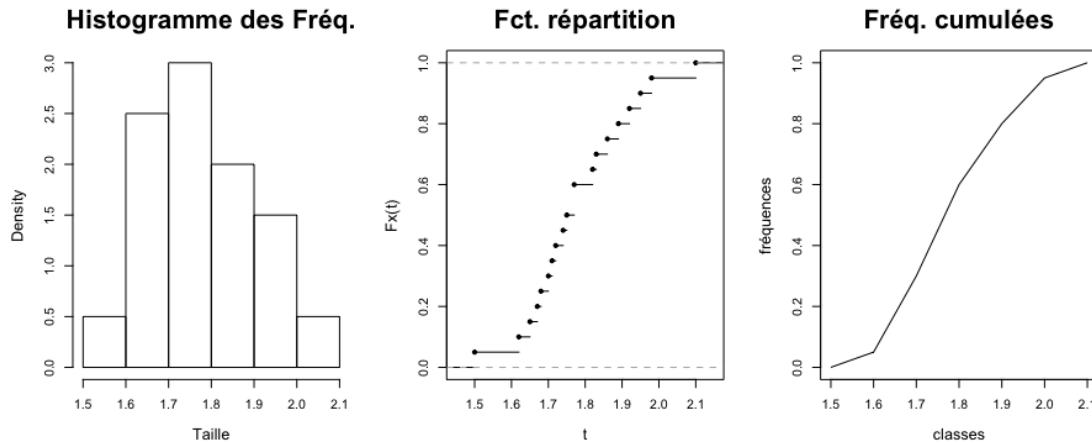
**Histogramme des Eff.**

```
> hist(Taille,freq=F,main="Histogramme des Fréq.",cex.main=2,cex.lab=1.2)

> plot(ecdf(Taille),main="Fct. répartition",xlab="t",ylab="Fx(t)",cex.main=2,cex.lab=1.2)

> tableau = data.frame(class=A$breaks,eff=c(0,A$counts),
+                         freq.cum=cumsum(c(0,A$counts))/sum(A$counts))
> tableau
  class eff freq.cum
1 1.5   0     0.00
2 1.6   1     0.05
3 1.7   5     0.30
4 1.8   6     0.60
5 1.9   4     0.80
6 2.0   3     0.95
7 2.1   1     1.00

> plot(tableau$class,tableau$freq.cum,type="l",main="Fréq. cumulées",
+       xlab="classes",ylab="fréquences",cex.main=2,cex.lab=1.2)
```



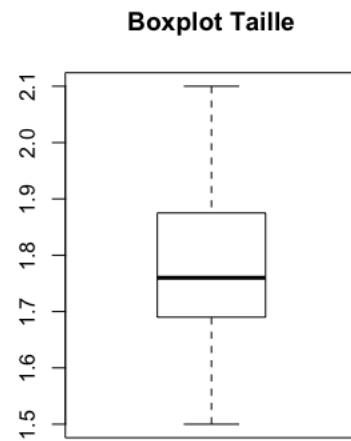
```
> summary(Taille)
   Min. 1st Qu. Median    Mean 3rd Qu.    Max.
1.500  1.695  1.760  1.782  1.867  2.100

> mean(Taille)
[1] 1.7815

> median(Taille)
[1] 1.76

> quantile(Taille)
 0%   25%   50%   75%  100%
1.5000 1.6950 1.7600 1.8675 2.1000

> boxplot(Taille,main="Boxplot Taille")
```



```
> n = length(Taille)                               > max(Taille) - min(Taille)
> s2=sum((Taille-mean(Taille))^2) / n ; s2        [1] 0.6
[1] 0.01870275                                     > IQR(Taille)
                                                       [1] 0.1725
> S2=s2 * n/(n-1) ; S2 #variance corrigée       > library(e1071)
[1] 0.01968711                                       > skewness(Taille)
                                                       [1] 0.3083229
> var(Taille)                                      > kurtosis(Taille)
[1] 0.01968711                                       [1] -0.3154956
> sd(Taille)
[1] 0.1403107
```

Q: Pourquoi 1,5 dans le boxplot ?

Dans le cas gaussien  $\mathcal{N}(0,1)$ ,

$$-q_{0.25} = q_{0.75} = 0.675$$

$$\text{d'où } \text{IC} = q_{0.75} - q_{0.25} = 2 \times q_{0.75} = 1.35$$

On en déduit que

$$\begin{aligned} -L_- &= L_+ = q_{0.75} + 1,5 \times 2 \times q_{0.75} \\ &= 4 \times q_{0.75} \\ &= 2.7 \end{aligned}$$

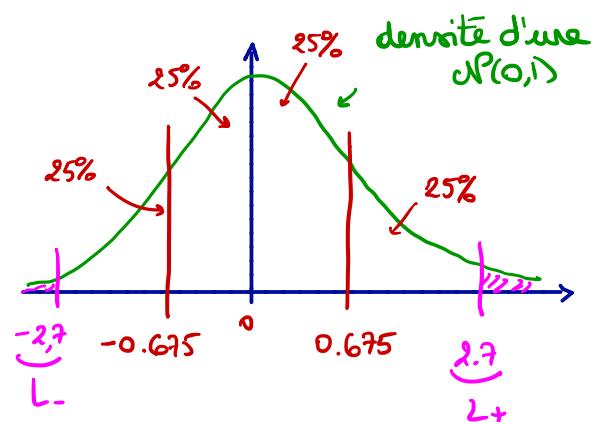
D'où, si  $Z \sim \mathcal{N}(0,1)$ ,

$$\mathbb{P}(L_- \leq Z \leq L_+) = 99,3\%$$

De la même manière, on obtient

pas assez arrondi  $\rightarrow$

$$\begin{aligned} (\text{Rappel: } L_+ &= q_{0.75} + 1,5 \text{ IC} \\ L_- &= q_{0.25} - 1,5 \text{ IC} ) \end{aligned}$$



$z$	$L_+ = -L_-$	$\mathbb{P}(L_- \leq z \leq L_+)$
1	2.02	95,7 %
1.5	2.7	99,3 %
1.583	2.81	99,5 %
2	3.37	99,9 %

trop d'outliers

✓

trop large

# Chapitre 3

## Statistique descriptive bidimensionnelle

On s'intéresse à présent à l'étude simultanée de deux variables  $X$  et  $Y$  observées sur un même échantillon ; il s'agit de l'*étude bidimensionnelle* (ou *bivariée*). Notons encore  $n$  la taille d'échantillon et

$$(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

les observations (chaque couple correspondant à un individu).

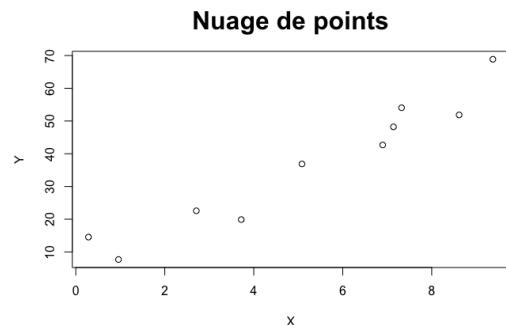
L'objectif de la statistique descriptive bidimensionnelle est de mettre en évidence d'éventuelles liaisons entre deux variables. Par exemple, pour un groupe d'individus, on peut penser qu'il existe une relation entre la taille et le poids. Comme pour l'étude unidimensionnelle, nous présenterons les représentations graphiques ainsi que les caractéristiques numériques exprimant cette liaison. Les notions introduites dépendant encore de la nature des variables, ce chapitre se décompose en trois parties.

### 3.1 Deux variables quantitatives

Supposons dans cette partie que  $X$  et  $Y$  sont deux variables quantitatives.

#### 3.1.1 Représentation graphique

Lors de l'étude de deux variables quantitatives, une représentation très commode est le *nuage de points* (*scatter plot* en anglais). Elle consiste à représenter chaque individu  $i$  par un point d'abscisse  $x_i$  et d'ordonnée  $y_i$ . L'ensemble de ces points (donnant son nom au graphique) donne en général une idée assez bonne de la variation conjointe des deux variables.



#### 3.1.2 Indices de liaison

**Covariance.** La *covariance*, parfois notée  $s_{xy}$ , est une généralisation bidimensionnelle de la variance. Elle est définie par

$$s_{xy} := \text{Cov}(\underline{x}, \underline{y}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}_n \bar{y}_n,$$

avec

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{et} \quad \bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i.$$

$$\text{appliquée à } \begin{cases} a_i = x_i - \bar{x}_n \\ b_i = y_i - \bar{y}_m \end{cases}$$

C'est donc la moyenne des produits des écarts aux moyennes.

La covariance vérifie les propriétés suivantes.

- Contrairement à la variance, la covariance peut prendre toute valeur réelle (et en particulier les valeurs négatives).
- La covariance entre une variable et elle-même n'est autre que sa variance :

$$s_{xx} = \text{Cov}(\underline{x}, \underline{x}) = \text{Var}(\underline{x}) = s_x^2.$$

- C'est un indice symétrique :  $\text{Cov}(\underline{x}, \underline{y}) = \text{Cov}(\underline{y}, \underline{x})$ .
- La covariance vérifie un propriété mathématique intéressante, appelée l'*inégalité de Cauchy-Schwarz*:

$$|\text{Cov}(\underline{x}, \underline{y})| \leq \sqrt{\text{Var}(\underline{x}) \text{Var}(\underline{y})}.$$

on obtient

$$|\text{cov}(\underline{x}, \underline{y})| \leq \sqrt{\text{var}(\underline{x}) \times \text{var}(\underline{y})}$$

**Corrélation.** La covariance dépendant des unités de mesure des deux variables considérées, on peut la rendre intrinsèque en renormalisant chaque variable (centrée) par son écart-type, ce qui revient à diviser la covariance par le produit des écart-types. On définit ainsi la corrélation linéaire par

$$r_{xy} := \text{Cor}(\underline{x}, \underline{y}) = \frac{\text{Cov}(\underline{x}, \underline{y})}{\sqrt{\text{Var}(\underline{x}) \text{Var}(\underline{y})}} = \frac{s_{xy}}{s_x s_y}.$$

Remarquons que par l'inégalité de Cauchy-Schwarz, le coefficient de corrélation linéaire vérifie  $r_{xy} \in [-1, 1]$ . Il mesure la dépendance linéaire entre deux variables et peut s'interpréter comme suit.

- Le signe de la corrélation détermine le sens de la liaison entre les deux variables :
  - si  $r_{xy} > 0$ , alors les deux variables auront tendance à varier dans le même sens : si  $X$  augmente,  $Y$  augmente également.
  - si  $r_{xy} < 0$ , alors les deux variables auront tendance à varier dans des sens opposés : si  $X$  augmente,  $Y$  diminue et inversement.
- La valeur absolue du coefficient indique l'intensité de la liaison :
  - si  $|r_{xy}|$  est proche de 1, les points sont alignés le long d'une droite.
  - si  $|r_{xy}|$  est proche de 0, la liaison est plutôt faible.

**Régression linéaire.** Lorsque l'on voit apparaître une liaison entre les deux variables, et que l'on peut considérer, a priori, que l'une (disons  $X$ ) est cause de l'autre (disons  $Y$ ), il est naturel de chercher une fonction  $f$  de  $X$  approchant  $Y$  "au mieux". La méthode statistique permettant de trouver une telle fonction s'appelle la *regression*.

Pour mettre en œuvre une regression, il est nécessaire de faire des hypothèses sur la forme de  $f$ . En particulier, lorsque le coefficient de correlation est proche de 1 en valeur absolue, il semblerait qu'il y ait une relation linéaire entre la variable  $X$  et la variable  $Y$ , menant à chercher une fonction de la forme

$$f(x) = a + bx.$$

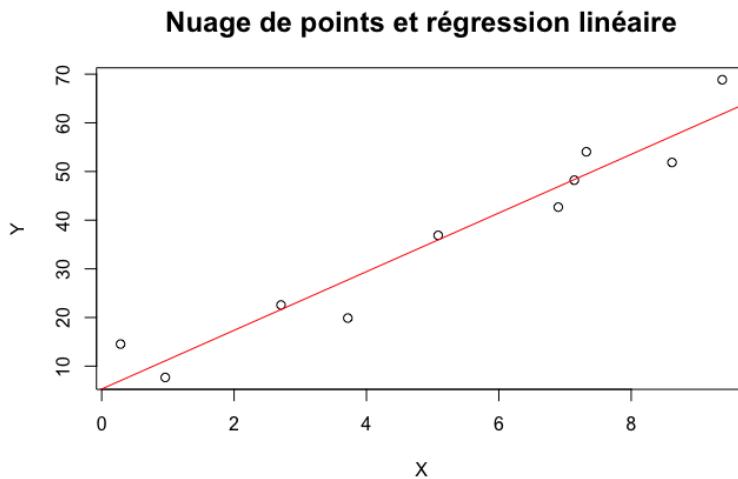
Afin de déterminer  $a$  et  $b$  pour que  $f$  explique "au mieux" la liaison entre  $X$  et  $Y$ , on définit le *critère des moindres carrés* par

$$h(a, b) = \sum_{i=1}^n (y_i - [a + bx_i])^2.$$

C'est la somme des carrés des distances entre la valeur observée  $y_i$  pour  $Y$  et son "explication" par  $f$  appliquée à la valeur observée  $x_i$  pour  $X$ . La minimisation de ce critère fournit l'unique solution

$$\hat{a} = \bar{y}_n - \hat{b}\bar{x}_n \quad \text{et} \quad \hat{b} = \frac{\text{Cov}(\underline{x}, \underline{y})}{\text{Var}(\underline{x})} = \frac{s_{xy}}{s_x^2}.$$

La droite d'équation  $y = \hat{a} + \hat{b}x$  est appelée *droite de régression* (représentée en continu ci-dessous).



Les valeurs  $\hat{y}_i = \hat{a} + \hat{b}x_i$  sont appelées les *valeurs prédictes* et les valeurs  $\hat{\epsilon}_i = y_i - \hat{y}_i$  sont appelés les *résidus*. L'étude théorique de la régression linéaire n'est pas détaillée dans ce cours et fait l'objet d'une Unité de Formation complète en 4ème année (GMM).

## 3.2 Une variable quantitative et une variable qualitative

Supposons dans cette partie que  $X$  est une variable qualitative, de modalités  $m_1, \dots, m_J$ , et que  $Y$  est une variable quantitative. Afin d'étudier d'éventuelles liaisons entre  $X$  et  $Y$ , l'idée est de regrouper les données en différentes classes :

$$\mathcal{C}_j = \{1 \leq i \leq n ; x_i = m_j\}.$$

Ainsi la  $j$ ème classe contient tous les individus pour lesquels la variable  $X$  prend la modalité  $j$ . Par exemple, si  $X$  est la variable "couleur des yeux" et  $Y$  est la variable "taille", nous aurons trois classes, regroupant respectivement les individus ayant les yeux bleus, ceux ayant les yeux verts et ceux ayant les yeux marrons.

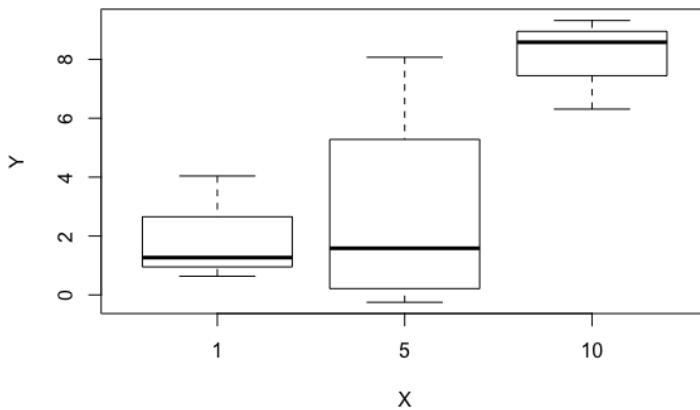
L'ensemble des classes  $\mathcal{C}_j$  définit une partition de l'ensemble des individus  $\{1, \dots, n\}$ . On note pour chaque  $j$  dans  $\{1, \dots, J\}$ ,  $n_j$  l'effectif de la classe  $\mathcal{C}_j$  (avec toujours  $\sum_{j=1}^J n_j = n$ ). On peut alors définir la moyenne et la variance de  $Y$  pour chaque classe :

$$\bar{y}_{[j]} = \frac{1}{n_j} \sum_{i \in \mathcal{C}_j} y_i, \quad \text{et} \quad s_{y,[j]}^2 = \frac{1}{n_j} \sum_{i \in \mathcal{C}_j} (y_i - \bar{y}_{[j]})^2.$$

### 3.2.1 Représentation graphique

Afin de représenter l'influence éventuelle de la variable  $X$  sur la variable  $Y$ , on trace les diagrammes en boîtes (boîtes à moustaches) pour chaque classe sur un même graphique : c'est le *diagramme en boîtes parallèles*.

### Boxplots parallèles

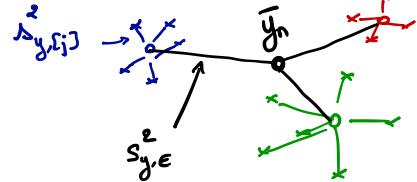


Plus les boîtes sont positionnées différemment, plus les variables  $X$  et  $Y$  sont liées.

#### 3.2.2 Indice de liaison

Avant de définir un indice de liaison entre les deux variables, regardons comment se décomposent la moyenne et la variance de  $Y$  sur la partition définie par les modalités de  $X$ . Pour la moyenne, on obtient la décomposition suivante :

$$\bar{y}_n = \frac{1}{n} \sum_{j=1}^J n_j \bar{y}_{[j]}.$$



La variance se décompose en une somme de deux termes :

$$s_y^2 = \underbrace{\frac{1}{n} \sum_{j=1}^J n_j (\bar{y}_{[j]} - \bar{y}_n)^2}_{s_{y,E}^2} + \underbrace{\frac{1}{n} \sum_{j=1}^J n_j s_{y,[j]}^2}_{s_{y,R}^2} = s_{y,E}^2 + s_{y,R}^2,$$

avec

- $s_{y,E}^2$  la *variance expliquée* (par  $X$ ), aussi appelée *variance inter-classes*. Elle représente ce que serait la variance de  $Y$  si, dans chaque classe de la partition,  $Y$  était constante (égale à  $\bar{y}_{[j]}$ , de sorte que  $s_{y,[j]}^2 = 0$ , et ce quelque soit  $j$ ).
- $s_{y,R}^2$  la *variance résiduelle*, aussi appelée *variance intra-classes*. Elle représente (en moyenne), ce qui reste de variabilité de  $Y$  à l'intérieur de chaque classe.

En effet,

$$\begin{aligned}
 s_y^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2 \\
 &= \frac{1}{n} \sum_{j=1}^J \sum_{i \in E_j} (y_i - \bar{y}_{E_j} + \bar{y}_{E_j} - \bar{y}_n)^2 \\
 &= \frac{1}{n} \sum_{j=1}^J \left[ \underbrace{\sum_{i \in E_j} (y_i - \bar{y}_{E_j})^2}_{m_j s_{y,[j]}^2} + \underbrace{\sum_{i \in E_j} (\bar{y}_{E_j} - \bar{y}_n)^2}_{m_j (\bar{y}_{E_j} - \bar{y}_n)^2} + 2 \underbrace{\left( \sum_{i \in E_j} (y_i - \bar{y}_{E_j}) \right)}_{\sum_{i \in E_j} y_i - m_j \bar{y}_{E_j}} \left( \bar{y}_{E_j} - \bar{y}_n \right) \right] \\
 &\quad \sum_{i \in E_j} y_i - m_j \bar{y}_{E_j} = 0 \\
 &= \frac{1}{n} \sum_{j=1}^J m_j s_{y,[j]}^2 + \frac{1}{n} \sum_{j=1}^J m_j (\bar{y}_{E_j} - \bar{y}_n)^2
 \end{aligned}$$

Intuitivement, plus  $s_{y,E}^2$  est grande par rapport à  $s_{y,R}^2$ , plus les variables  $X$  et  $Y$  sont liées. On introduit donc le *rappor de corrélation* entre les deux variables :

$$c_{y|x} = \sqrt{\frac{s_{y,E}^2}{s_y^2}}.$$

Attention,  $c_{y|x}$  n'est pas symétrique. En effet,  $X$  et  $Y$  n'étant de même nature, on ne peut pas les échanger.

D'après la formule de décomposition de la variance, on a

$$1 = c_{y|x}^2 + \frac{s_{y,R}^2}{s_y^2}.$$

Il en découle les propriétés suivantes :

- Le rapport de corrélations vérifie toujours  $0 \leq c_{y|x} \leq 1$ .
- Si  $c_{y|x} = 1$ , c'est-à-dire que  $s_{y,R}^2 = 0$ , on a pour tout  $j$ ,  $s_{y,[j]}^2 = 0$  i.e.  $Y$  est constante chacune des classes. Dans ce cas, la connaissance de  $X$  permet de déterminer complètement  $Y$  : il y a une liaison totale entre  $X$  et  $Y$ .
- Si au contraire,  $c_{y|x} = 0$ , c'est-à-dire que  $s_{y,E}^2 = 0$ , on a pour tout  $j$ ,  $\bar{y}_{[j]} = \bar{y}_n$ . Dans ce cas, en moyenne, la variable  $X$  n'a aucune influence sur  $Y$  : il n'y a pas de liaison entre  $X$  et  $Y$ .

### 3.3 Deux variables qualitatives

Dans cette partie, nous étudions le cas où  $X$  et  $Y$  sont deux variables qualitatives de modalités  $m_1, \dots, m_J$  et  $l_1, \dots, l_K$ . Il est commode de présenter les données sous forme de tableau de contingence (*contingency table* en anglais). Pour cela, on définit :

- les *effectifs conjoints*, comptant le nombre de fois que les modalités  $m_j$  et  $l_k$  sont observées sur un même individu,

$$n_{jk} = \sum_{i=1}^n \mathbb{1}_{\{x_i=m_j\}} \mathbb{1}_{\{y_i=l_k\}},$$

- les *effectifs marginaux*, comptant le nombre de fois que la variable  $X$  (respectivement  $Y$ ) prend la modalité  $m_j$  (respectivement  $l_k$ ) :

$$n_{j\cdot} = \sum_{k=1}^K n_{jk} \quad \text{et} \quad n_{\cdot k} = \sum_{j=1}^J n_{jk},$$

$X \backslash Y$	$l_1$	$\dots$	$l_k$	$\dots$	$l_K$	total
$m_1$	$n_{11}$	$\dots$	$n_{1k}$	$\dots$	$n_{1K}$	$n_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$m_j$	$n_{j1}$	$\dots$	$n_{jk}$	$\dots$	$n_{jK}$	$n_{j\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$m_J$	$n_{J1}$	$\dots$	$n_{Jk}$	$\dots$	$n_{JK}$	$n_{J\cdot}$
total	$n_{\cdot 1}$	$\dots$	$n_{\cdot k}$	$\dots$	$n_{\cdot K}$	$n$

Remarquons qu'il est possible de définir les notions de fréquences conjointes et les fréquence marginales de façon analogue.

### 3.3.1 Représentation graphique

Pour représenter graphiquement les liaisons entre la variable  $X$  et la variable  $Y$ , nous introduisons des quantités adaptées appelées les *profils*.

- Le  $j$ ème *profil-ligne* est l'ensemble des fréquences des modalités de la variable  $Y$  au sein de la classe  $C_j$  (*i.e.* conditionnelles au fait que  $X$  soit égale à la modalité  $m_j$ ) :

$$\left( \frac{n_{j1}}{n_{j\cdot}}, \dots, \frac{n_{jk}}{n_{j\cdot}}, \dots, \frac{n_{jK}}{n_{j\cdot}} \right) \in [0, 1]^K.$$

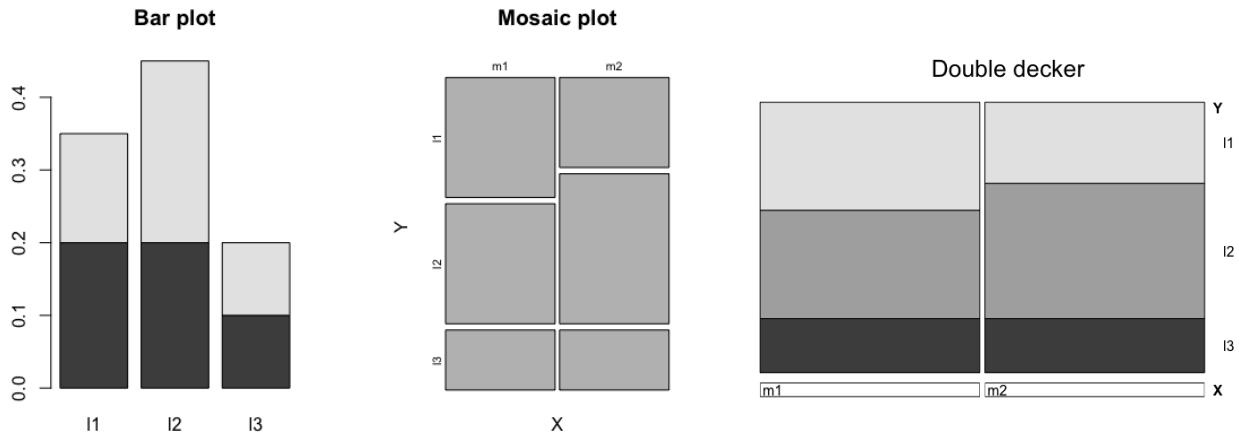
- De manière analogue, le  $k$ ème *profil-colonne* est l'ensemble des fréquences des modalités de la variable  $X$  conditionnellement au fait que  $Y$  soit égal à la modalité  $l_k$  :

$$\left( \frac{n_{1k}}{n_{\cdot k}}, \dots, \frac{n_{jk}}{n_{\cdot k}}, \dots, \frac{n_{Jk}}{n_{\cdot k}} \right) \in [0, 1]^J.$$

Remarquons que les renormalisations par les effectifs de la variable par laquelle on conditionne permet d'avoir des vecteurs qui se somment à 1, les rendant ainsi comparables (quelques soient les effectifs).

La représentation graphique des profils-lignes ou des profils-colonnes grâce à des diagrammes en barre (par exemple) donne une idée assez précise de la variation conjointe des deux variables.

Plusieurs fonctions  permettent de faire ces représentations.



### 3.3.2 Indices de liaison

Commençons par remarquer que les trois propriétés suivantes sont équivalentes :

- (i) Tous les profils-ligne sont égaux.
- (ii) Tous les profils-colonne sont égaux.
- (iii)  $\forall j \in \{1, \dots, J\}$  et  $\forall k \in \{1, \dots, K\}$ ,  $n_{jk} = \frac{n_{j\cdot} \times n_{\cdot k}}{n}$ .

*c.f. p. 30*

Dans ce cas, on dit qu'il n'existe aucune forme de liaison entre les deux variables considérées  $X$  et  $Y$ . On mesure ainsi la liaison en évaluant l'écart entre la situation observée et l'état de non liaison défini ci-dessus.

L'indice de liaison du *Khi-deux* (*Chi-square* en anglais) est défini par

$$\chi^2 = \sum_{j,k} \frac{\left( n_{jk} - \frac{n_{j \cdot} \cdot n_{\cdot k}}{n} \right)^2}{\frac{n_{j \cdot} \cdot n_{\cdot k}}{n}}.$$

Comme précisé ci-dessus, lorsque  $\chi^2 = 0$ , on en déduit qu'il n'y a pas de liaison entre les deux variables.

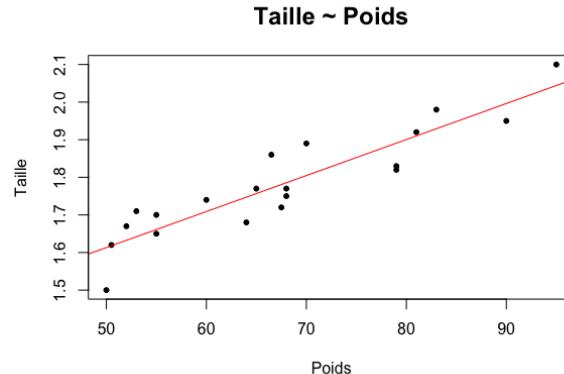
Un des inconvénients de cet indice est qu'il dépend de  $n$ , et ne permet donc pas de comparer des échantillons de taille différentes. Afin de pallier ce désavantage, on peut donc plutôt s'intéresser au *coefficient d'association de Pearson* (également appelé *phi-deux*) défini par

$$\Phi^2 = \frac{\chi^2}{n}.$$

## 3.4 Exemples avec R

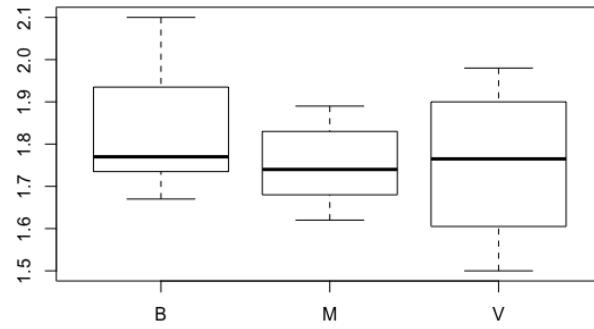
### VARIABLES "TAILLE" ET "POIDS"

```
> Taille<-Data[,"Taille"]          > coef(mod)
> Poids<-Data[,"Poids"]           (Intercept)      Poids
                                         1.134425853  0.009575644
                                         [1] 1.708566
                                         > plot(Poids,Taille,main="Taille ~ Poids")
                                         > abline(mod,col="red")
> cor(Taille,Poids)
[1] 0.9116097
> mod<-lm(Taille~Poids,data=Data)
> bhat = cov(Poids,Taille)/var(Poids)
> bhat
[1] 0.009575644
> ahat = mean(Taille) - bhat * mean(Poids)
> ahat
[1] 1.134426
```



### VARIABLES "TAILLE" ET "YEUX"

```
> Yeux<-Data[,"Couleur.des.yeux"]
> Taille<-Data[,"Taille"]
> boxplot(Taille~Yeux)
> library(BioStatR)
> eta2(Taille,Yeux)
[1] 0.09862483
```



## VARIABLES "YEUX" ET "SEXES"

```

> Yeux<-Data[,1]

> Sexe<-Data[,2]

> table.cont<-table(Yeux,Sexe)

> addmargins(table.cont)
  Sexe
Yeux   F   H Sum
  B    4   3   7
  M    4   5   9
  V    2   2   4
Sum 10 10 20

> prop.cont <- prop.table(table.cont)

> prop.cont
  Sexe
Yeux   F   H
  B 0.20 0.15
  M 0.20 0.25
  V 0.10 0.10

> library(vcd)

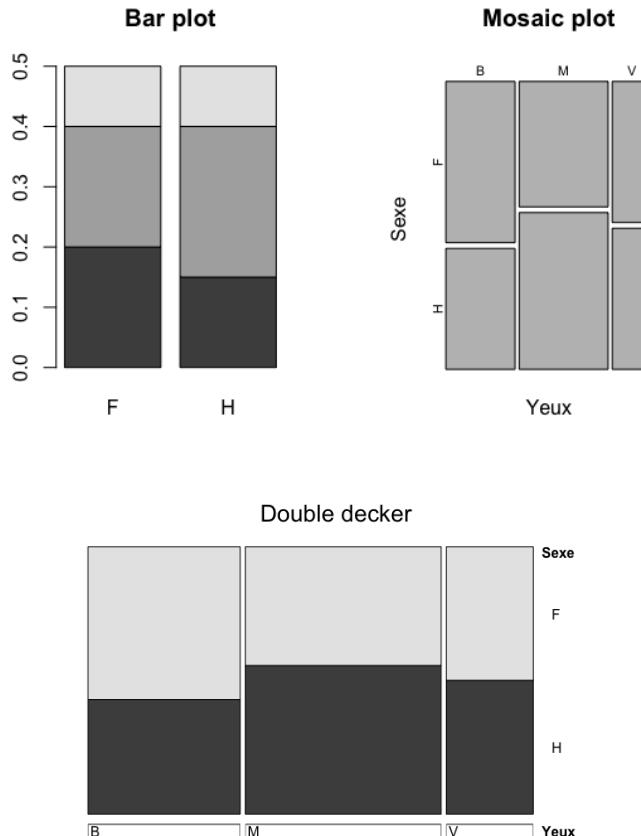
> par(mfrow=c(1,2))

> barplot(prop.cont,main="Bar plot")

> mosaicplot(table.cont,main="Mosaic plot")

> doubledecker(table.cont)

```



⊕ • Montreons que (i)  $\Rightarrow$  (iii). Rappel:  $\left( \frac{m_{j,k}}{m_{j,:}}, \dots, \frac{m_{j,k}}{m_{j,:}}, \dots, \frac{m_{j,k}}{m_{j,:}} \right)$  j<sup>e</sup> profil ligne.

Supposons (i). alors  $\forall j, \frac{m_{j,k}}{m_{j,:}} = p_k \in [0,1]$  (ne dépend pas de j)

$$\text{i.e. } m_{j,k} = p_k \times m_{j,:}$$

$$\text{Par ailleurs } m_{:,k} = \sum_{j=1}^J m_{j,k} = p_k \underbrace{\sum_{j=1}^J m_{j,:}}_n \quad \text{d'où } p_k = \frac{m_{:,k}}{n} \quad \text{i.e. } \frac{m_{j,k}}{m_{j,:}} = \frac{m_{:,k}}{n}$$

$$\text{i.e. } m_{j,k} = \frac{m_{:,k} \times m_{j,:}}{n}.$$

• Montreons que (iii)  $\Rightarrow$  (i)

$$\text{Si } m_{j,k} = \frac{m_{j,:} \cdot m_{:,k}}{n} \quad \text{alors} \quad \frac{m_{j,k}}{m_{j,:}} = \frac{m_{:,k}}{n} \quad \text{ne dépend pas de j. d'où (i)}$$

# Chapitre 4

## Analyse en Composantes Principales

### 4.1 Introduction

#### 4.1.1 Positionnement du problème

Dans les chapitres précédents, nous nous sommes intéressés à la description de variables en dimensions 1 et 2. Cependant, dans de nombreuses situations, le nombre de variables d'intérêt pour un problème donné peut être potentiellement très élevé, d'une dizaine d'unités à des milliers d'éléments. Parmi les domaines impliqués par de telles problématiques, on pourra citer entre autres

- **Marketing** : pour chaque client, on dispose d'un certain nombre de variables mesurant les pratiques en terme d'achat (en euros ou en volume) pour un certain nombre de catégories de produits : jardinage, consommables pour bébé, matériel électronique, alimentaire, etc...
- **Assurance** : on dispose, pour chaque assuré, de diverses informations parmi lesquelles l'ancienneté en tant qu'assuré, le nombre de sinistres, le montant des biens assurés, etc...
- **Analyses cliniques** : lors d'une étude clinique, les praticiens vont regrouper un certain nombre d'informations sur le panel de patients retenus : âge, poids, taille, nombre d'opérations, taux de glycémie, cholestérol, etc...

Plus spécifiquement, le jeu de données présenté dans la Table 4.1 permettra d'illustrer les différentes discussions proposées dans les pages suivantes. Ce dernier rassemble les notes d'un groupe fictif d'étudiants dans 4 matières.

D'un point de vue général, au fur et à mesure que le nombre de variables (et dans une moindre mesure d'individus) augmente, il devient de plus en plus difficile d'avoir une vision synthétique de la base de données considérée. Dans le même temps, il peut s'avérer primordial de pouvoir dégager de grandes tendances sur les "individus" concernés : identification de clients types en marketing ou de profils de patients remarquables dans des essais cliniques. Il devient dès lors nécessaire de se doter d'outils permettant la mise en place de ce type d'analyse. Dans le cadre du cours, nous allons aborder l'analyse en composantes principales (ACP) qui est utilisée lorsque les variables sont quantitatives. D'autres outils seront étudiés en 4ème année (GMM).

	Mathématiques	Physique	Français	Anglais
Nathan	6	6	5	5.5
Emma	8	8	8	8
Lola	6	7	11	9.5
Baptiste	14,5	14,5	15,5	15
Mathilde	14	14	12	12.5
Ines	11	10	5.5	7
Lucas	5.5	7	14	11.5
Aymeric	13	12.5	8.5	9.5
Chloe	9	9.5	12.5	12

TABLE 4.1 – Notes d'un "panel" d'étudiants dans quatre matières différentes.

#### 4.1.2 Transformation des données

D'un point de vue formel, on dispose initialement d'un tableau de données représenté par une matrice  $X = (X_{ij})_{i=1..n, j=1..p}$ ,  $p$  et  $n$  désignant respectivement le nombre de variables quantitatives et d'individus. Autrement dit, on travaille avec

$$X = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \vdots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix},$$

où pour tout  $i \in \{1, \dots, n\}$ ,  $j \in \{1, \dots, p\}$ ,

- la  $i^{\text{ème}}$  ligne  $X_i := (X_{i1}, \dots, X_{ip})$  désigne les observations relatives à l'**individu**  $i$ ,
- la  $j^{\text{ème}}$  colonne  $X^{(j)} := (X_{1j}, \dots, X_{nj})'$  rassemble les informations relatives à la **variable**  $j$ .

Il est courant en pratique de construire une ACP à partir d'une transformation de la matrice initiale  $X$ . Par la suite, la matrice  $T$  désignera la matrice de "travail", cette dernière pouvant prendre différentes formes. Le choix de  $T$  conditionne implicitement l'individu (fictif) de référence  $\Omega$ , auquel seront comparés les autres individus. Différentes options sont usuellement envisagées :

- $T = X$  : on travaille dans ce cas avec les données brutes. Implicitement, l'individu de référence  $\Omega$  est le vecteur nul de dimension  $p$  (chacune des variables de cet individu  $\Omega$  prend la valeur 0).
- $T = (X^{(1)} - \bar{X}^{(1)}, \dots, X^{(p)} - \bar{X}^{(p)})$  où  $\bar{X} = (\bar{X}^{(1)}, \dots, \bar{X}^{(p)})$  avec  $\bar{X}^{(j)} := n^{-1} \sum_{i=1}^n X_{ij}$  pour tout  $j \in \{1, \dots, p\}$ . Dans cette situation, il est possible de voir que chaque colonne de la matrice  $T$  est centrée<sup>1</sup>. On parlera dans ce cadre d'**ACP centrée**. L'individu de référence  $\Omega$  sera ici l'individu "moyen", i.e. celui pour lequel chacune des variables prendra la valeur moyenne de celle de l'échantillon.

- on peut également "réduire" les données en posant  $T = \left( \frac{X^{(1)} - \bar{X}^{(1)}}{s^{(1)}}, \dots, \frac{X^{(p)} - \bar{X}^{(p)}}{s^{(p)}} \right)$ , avec

$$s^{(j)} := \sqrt{\frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}^{(j)})^2}, \quad \forall j \in \{1, \dots, p\}. \quad (4.1)$$

1. La soustraction ci-dessus est effectuée coordonnée par coordonnée

Travailler avec ce type de matrice conduira à mettre en place une **ACP centrée réduite**. L'individu moyen reste l'individu de référence, mais les "écart" sont normalisés. Ce type de configuration est en particulier à privilégier lorsqu'on regroupe des variables exprimées dans des unités différentes, ou qui ne sont pas du tout à la même échelle.

Une fois le choix de  $T$  effectué, il reste à se doter d'outils permettant de mesurer à la fois les distances entre individus et entre variables. Cette étape sera détaillée dans les sections suivantes.

### En pratique

Revenons au jeu de données présenté dans la Table 4.1. La matrice  $X$  est donnée par

$$X = \begin{pmatrix} 6 & 6 & 5 & 5.5 \\ 8 & 8 & 8 & 8 \\ 6 & 7 & 11 & 9.5 \\ 14.5 & 14.5 & 15.5 & 15 \\ 14 & 14 & 12 & 12.5 \\ 11 & 10 & 5.5 & 7 \\ 5.5 & 7 & 14 & 11.5 \\ 13 & 12.5 & 8.5 & 9.5 \\ 9 & 9.5 & 12.5 & 12 \end{pmatrix}, \quad \text{avec } n = 9 \text{ et } p = 4.$$

Chaque colonne (dans  $\mathbb{R}^9$ ) donne les notes d'une matière, alors que chaque ligne représente les notes d'un individu (voir la Table 4.1 pour la correspondance). Conserver la matrice telle quelle a peu de sens dans la mesure où l'individu de "référence" associé au vecteur nul n'est pas vraiment représentatif de la population. Dans la mesure où toutes les variables correspondent à des notes, il ne semble pas nécessaire de procéder à une renormalisation des données. Nous travaillerons par la suite avec la matrice centrée.

Remarquons dans un premier temps que

$$\bar{X} = (\bar{X}^{(1)}, \dots, \bar{X}^{(p)}) = (9.67, 9.83, 10.22, 10.06).$$

Cela correspond aux moyennes dans chaque matière (i.e. de chaque variables). La matrice  $T$  recentrée est alors égale à

$$T = \begin{pmatrix} -3.67 & -3.83 & -5.22 & -4.56 \\ -1.67 & -1.83 & -2.22 & -2.06 \\ -3.67 & -2.83 & 0.78 & -0.56 \\ 4.83 & 4.67 & 5.28 & 4.94 \\ 4.33 & 4.17 & 1.78 & 2.44 \\ 1.33 & 0.17 & -4.72 & -3.06 \\ -4.17 & -2.83 & 3.78 & 1.44 \\ 3.33 & 2.67 & -1.72 & -0.56 \\ -0.67 & -0.33 & 2.28 & 1.94 \end{pmatrix}. \quad (4.2)$$

Dans ce contexte, chaque colonne représente les écarts à la moyenne dans la matière correspondante.

## 4.2 Inertie

### 4.2.1 Représentation des individus et des variables

Le tableau  $T$  peut être étudié au premier abord de deux manières possibles : soit à travers l'étude des individus, soit par l'intermédiaire de l'analyse des variables. Nous verrons plus loin que l'analyse en composantes principales permet une prise en compte simultanée de ces deux points de vue. Dans tous les cas, il nous faut dès à présent définir formellement les deux espaces de travail considérés.

#### a) Espace des individus

Chaque individu de la matrice  $T$  est un élément de  $\mathbb{R}^p$ . Afin de mesurer la distance entre deux individus, il est donc nécessaire de munir  $\mathbb{R}^p$  d'une norme, et donc d'un produit scalaire. On se donne pour cela une matrice  $M$  diagonale  $p \times p$  définie comme

$$M = \begin{pmatrix} m_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & m_p \end{pmatrix},$$

où  $(m_1, \dots, m_p)$  est une suite strictement positive définie par l'utilisateur. On munit alors  $\mathbb{R}^p$  du produit scalaire  $\langle \cdot, \cdot \rangle_M$  où pour tout  $a, b \in \mathbb{R}^p$ ,

$$\begin{aligned} a &= \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} \\ b &= \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} \end{aligned}$$

$$\langle a, b \rangle_M := a' M b = \sum_{j=1}^p m_j a_j b_j. \quad \begin{aligned} a' M b &= (a_1 \dots a_p) \begin{pmatrix} m_1 & & 0 \\ 0 & \ddots & \\ & & m_p \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} \\ &= (a_1 \dots a_p) \begin{pmatrix} m_1 b_1 \\ \vdots \\ m_p b_p \end{pmatrix} \end{aligned}$$

On a alors en particulier

$$\|a\|_M^2 = \sum_{j=1}^p m_j a_j^2 \quad \forall a \in \mathbb{R}^p. \quad = \sum_{j=1}^p a_j m_j b_j$$

Bien qu'il n'y ait pas de restriction particulière concernant le choix de la suite  $(m_1, \dots, m_p)$ , on distingue en pratique deux choix possibles :

- la métrique identité :  $m = (1, \dots, 1)$ . Dans ce cas,  $\langle \cdot, \cdot \rangle$  correspond au produit scalaire euclidien dans  $\mathbb{R}^p$ . La distance entre deux individus est alors mesurée par l'intermédiaire de la norme euclidienne de la différence des coordonnées.
- pondération par les variances :

$$m = \left( \frac{1}{(s^{(1)})^2}, \dots, \frac{1}{(s^{(p)})^2} \right),$$

où les  $s^{(j)}$  sont les écart-types définis en (4.1). Utiliser cette métrique revient à pondérer la différence entre chaque coordonnée par l'inverse de l'écart-type de chaque variable. Cette métrique permet en particulier de réduire l'importance des variables possédant une trop "forte" variabilité.

En effet, *Mentionnons que considérer  $M = \text{diag}(m_1, \dots, m_p)$  pour  $T$  revient à transformer  $T$  en  $\tilde{T} = (\sqrt{m_1} T^{(1)}, \dots, \sqrt{m_p} T^{(p)}) \in \mathcal{M}_{n,p}(\mathbb{R})$  et considérer  $\tilde{M} = I_p$ .*

$$\begin{aligned} \text{Soient } i, k \in \{1, \dots, n\}, \quad \langle \tilde{T}_i, \tilde{T}_k \rangle_{\tilde{M}} &= \sum_{j=1}^p \tilde{T}_{ij} \tilde{T}_{kj} = \sum_{j=1}^p \left[ \sqrt{m_j} T_{ij} \times \sqrt{m_j} T_{kj} \right] \\ &= \sum_{j=1}^p m_j T_{ij} T_{kj} = \langle T_i, T_k \rangle_M \end{aligned}$$

Ici, lorsque  $M = \text{diag}\left(\frac{1}{(s^{(1)})^2}, \dots, \frac{1}{(s^{(p)})^2}\right)$  et  $T$  le tableau centré, cela revient à travailler avec le tableau  $\tilde{T}$  centré-réduit et la métrique  $\tilde{M} = I_p$ .

### b) Espace des variables

Chaque variable (colonne) du tableau  $T$  est un vecteur de taille  $n$  et appartient donc à  $\mathbb{R}^n$ . Soit  $W$  une matrice diagonale  $n \times n$  définie comme

$$W = \begin{pmatrix} w_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_n \end{pmatrix},$$

pour une suite  $w = (w_1, \dots, w_n)$  prédéfinie. La matrice  $W$  est souvent appelée *matrice des pondérations* dans la mesure où elle conditionne, d'une certaine manière, le poids accordé à chaque individu.

Sauf cas très particulier, on travaille dans la grande majorité des cas rencontrés avec la matrice

$$W = \frac{1}{n} I_n,$$

où  $I_n$  désigne la matrice identité dans  $\mathbb{R}^n$ . Pour tout  $u, v \in \mathbb{R}^n$ , on définit alors

$$\langle u, v \rangle_W := \frac{1}{n} \sum_{i=1}^n u_i v_i \quad \text{et} \quad \|u\|_W^2 = \frac{1}{n} \sum_{i=1}^n u_i^2.$$

#### 4.2.2 Inertie globale

On se place à présent dans l'espace des individus. Dans ce cas, le tableau  $T$  nous donne un nuage de  $n$  points, chaque point étant un élément de  $\mathbb{R}^p$ . La notion d'inertie suivante permet de donner une mesure de la dispersion de ce nuage par rapport à l'individu de référence  $\Omega$ .

##### Définition 4.1.

L'inertie du nuage de points relative à l'individu de référence  $\Omega$  est définie comme

$$I_\Omega := \sum_{i=1}^n w_i \|T_i\|_M^2. \tag{4.3}$$

Pour tout  $i \in \{1, \dots, n\}$ ,  $\|T_i\|_M^2$  représente la distance entre l'individu  $i$  et l'individu de référence. L'inertie globale  $I_\Omega$  correspond alors à la moyenne de ces distances. Typiquement, si l'inertie est "grande", cela signifie que la distance moyenne entre chaque individu et l'individu de référence  $\Omega$  est importante : le nuage est "très" dispersé. Au contraire, un nuage de point relativement "ramassé" autour de l'individu de référence sera associé à une faible inertie.

La proposition suivante donne une formule alternative pour l'inertie. On définit pour cela la matrice  $\Gamma$  comme

$$\Gamma = T'WT. \tag{4.4}$$

##### Proposition 4.1.

*L'inertie  $I_\Omega$  introduite en (4.3) peut s'écrire sous la forme*

$$I_\Omega = \text{Tr}(\Gamma M),$$

*où  $\text{Tr}(A)$  désigne la trace d'une matrice carrée  $A$ .*

$$\Gamma M = T' W T M \in \mathcal{M}_p(\mathbb{R})$$

p × n    ↓    n × p    ↓  
 n × n         p × p

multiplier à droite par une matrice diagonale revient à multiplier les colonnes.

## 4.2 INERTIE

**PREUVE :**  $\text{Tr}(\Gamma M) = \sum_{j=1}^p (\Gamma M)_{jj}$

avec  $(\Gamma M)_{jj} = \Gamma_{jj} m_j$

$$= (T' W T)_{jj} m_j$$

$$= \sum_{i=1}^n T_{ij} w_i T_{ij} m_j = \sum_{i=1}^n w_i m_j T_{ij}^2$$

ainsi  $\text{Tr}(\Gamma M) = \sum_{j=1}^p \sum_{i=1}^n w_i m_j T_{ij}^2$

$$= \sum_{i=1}^n w_i \underbrace{\left( \sum_{j=1}^p m_j T_{ij}^2 \right)}_{\|T_i\|_M^2} = \sum_{i=1}^n w_i \|T_i\|_M^2 = I_{\Omega}.$$

□

Nous verrons en TD que dans le cadre standard ( $M = I_p$  et  $W = n^{-1}I_n$ ), cela revient d'une certaine manière à mesurer la variabilité dans un cadre multidimensionnel. En effet,

- Dans le cas où  $T$  correspond au tableau centré,  $\Gamma$  correspond exactement à la matrice de variance-covariance de  $X$  (ou de  $T$ ). En particulier, les éléments diagonaux correspondent aux variances de chacune des variables d'intérêt, et on obtient donc

$$I_{\Omega} = \text{Tr}(\Gamma M) = \text{Tr}(\Gamma) = \sum_{j=1}^p (s^{(j)})^2.$$

- Dans le cas où  $T$  correspond au tableau centré et réduit,  $\Gamma$  correspond exactement à la matrice des corrélations de  $X$  (ou de  $T$ ), et dans ce cas

$$I_{\Omega} = \text{Tr}(\Gamma M) = \text{Tr}(\Gamma) = p.$$

### En pratique

On travaille à nouveau avec la matrice  $T$  centrée, définie en (4.2). En posant  $W_n = n^{-1}I_n$  et  $M = I_p$ , on obtient

$$\Gamma = \begin{pmatrix} 11.39 & 9.92 & 2.66 & 4.82 \\ 9.92 & 8.94 & 4.12 & 5.48 \\ 2.66 & 4.12 & 12.06 & 9.29 \\ 4.82 & 5.48 & 9.29 & 7.91 \end{pmatrix}.$$

On peut au passage vérifier que les éléments diagonaux correspondent bien à la variance de chaque matière. Dans ce cadre, l'inertie vaut alors

$$I_{\Omega} = 11.39 + 8.95 + 12.06 + 7.91 = 40.3.$$

Cette quantité mesure la dispersion du nuage de points, i.e. l'hétérogénéité du groupe d'étudiants sur l'ensemble des quatre matières.

#### 4.2.3 Inertie axiale

Dans la section précédente, la notion d'inertie globale permettait de mesurer la dispersion (variabilité) du nuage de points dans son ensemble. On va ici s'intéresser à la notion de dispersion sur un axe  $a$  donné.

On rappelle que l'on travaille toujours dans l'espace des individus. Chaque point de cet espace est un élément de  $\mathbb{R}^p$ . Soit  $a$  un vecteur  $M$ -normé de  $\mathbb{R}^p$ , i.e. tel que

$$\|a\|_M^2 = a'Ma = 1.$$

Etant donné un point  $x \in \mathbb{R}^p$ , la projection orthogonale de  $x$  sur la droite engendrée par  $a$  est égale à

$$P_a(x) = (x'Ma)a = \langle x, a \rangle_M a.$$

Etant donné un axe<sup>2</sup>  $a$ , on peut s'intéresser à la dispersion de la projection  $M$ -orthogonale du nuage de points sur cet axe. C'est le sens de la définition suivante.

#### Définition 4.2.

Soit  $a \in \mathbb{R}^p$   $M$ -normé. L'inertie sur l'axe  $a$ ,  $I_\Omega(a)$ , est définie comme

$$I_\Omega(a) = \sum_{i=1}^n w_i \frac{\langle T_i, a \rangle_M^2}{\|P_a(T_i)\|_M^2}.$$

En particulier,

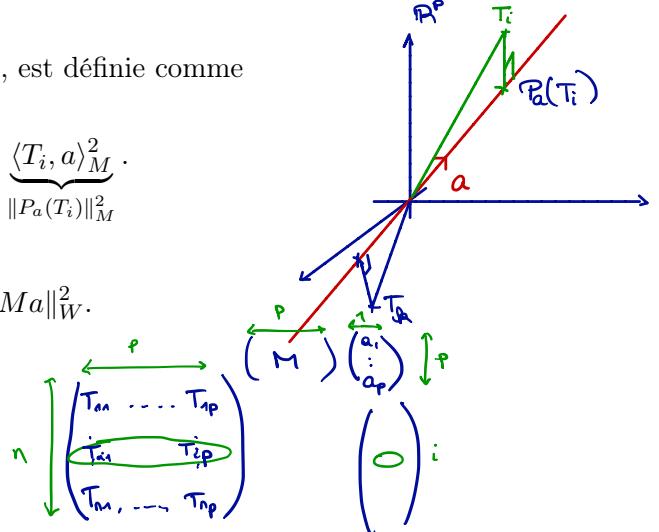
$$I_\Omega(a) = \|T M a\|_W^2.$$

En effet,  $\|T M a\|_W^2 = \sum_{i=1}^n w_i (\langle T M a \rangle_i^2)$

$$(\langle T M a \rangle_i)_i = T_i M a = \langle T_i, a \rangle_M$$

$$\text{d'où } \|T M a\|_W^2 = \sum_{i=1}^n w_i \langle T_i, a \rangle_M^2 = I_\Omega(a).$$

Remarque:  $T M a = \begin{pmatrix} \langle T_1, a \rangle_M \\ \vdots \\ \langle T_n, a \rangle_M \end{pmatrix} \in \mathbb{R}^n$



La matrice  $T$  étant de taille  $n \times p$ , le vecteur  $T M a$  est de taille  $n$ , dont chaque entrée  $i$  correspond à la projection de l'individu  $i$  sur l'axe  $a$ .

#### Proposition 4.2.

Soit  $(e_j)_{1 \leq j \leq p}$  une base  $M$ -orthonormée de  $\mathbb{R}^p$ . Alors  $I_\Omega = \sum_{j=1}^p I_\Omega(e_j)$ .

PREUVE :

$$\begin{aligned} \sum_{j=1}^p I_\Omega(e_j) &= \sum_{j=1}^p \sum_{i=1}^n w_i \langle T_i, e_j \rangle_M^2 \\ &= \sum_{i=1}^n w_i \underbrace{\sum_{j=1}^p \langle T_i, e_j \rangle_M^2}_{} \\ &= \sum_{i=1}^n w_i \|T_i\|_M^2 \quad \text{car } (e_1, \dots, e_p) \text{ est une base } M - \text{L.I. de } \mathbb{R}^p \\ &= I_{\Omega_M}. \end{aligned}$$

$$\text{d'où } T_i = \sum_{j=1}^p \langle T_i, e_j \rangle_M e_j$$

$$\text{et on a bien } \|T_i\|_M^2 = \sum_{j=1}^p \langle T_i, e_j \rangle_M^2.$$

□

2. Par abus de notation, on confondra souvent l'axe  $a$  et la droite engendrée par  $a$

Ainsi, on peut voir que l'inertie globale peut être décomposée comme la somme des inerties sur chacun des axes de la base. Autrement dit, la dispersion globale du nuage de points peut être vue comme la somme des dispersions dans chacune des directions.

### En pratique

Les quatre variables de la Table 4.1 sont respectivement associées aux inerties axiales 11.39, 8.94, 12.06 et 7.91. Sachant que l'inertie globale est égale pour ce jeu de données à 40.3, il est possible de calculer la part d'inertie (en %) apportée par chacune des matières :

Matière	Pourcentage d'inertie
Mathématiques	28.3 %
Physique	22.2 %
Français	29.9 %
Anglais	19.6 %

En effet, dans le cas  $M = I_p$ , la base canonique est une base  $M$ -orthonormée.

## 4.3 Recherche des composantes principales

L'analyse en composantes principales vise à produire une description synthétique d'un tableau de données relatif à  $p$  variables et  $n$  individus. Dans un certain nombre de situations, il est cependant raisonnable d'imaginer que certaines informations contenues dans les différentes variables du tableau sont "redondantes" (i.e. certaines variables sont corrélées). L'objectif de l'ACP est d'exhiber un petit nombre de **meta-variables** (directions) permettant de rendre compte de la dispersion des données d'intérêt.

### 4.3.1 Première composante principale

On va se concentrer dans un premier temps sur la définition et la recherche de la première composante principale. On a vu dans la section précédente que chaque axe canonique permettait d'expliquer une (petite) partie de la variabilité totale d'un nuage de points. Plutôt que de se contenter de travailler avec cette base canonique, on va chercher à proposer un changement de repère associé à une décomposition plus parcimonieuse de l'inertie. Ce principe donne lieu à la définition suivante

#### Définition 4.3.

Le premier axe principal  $a_1$  est l'axe maximisant  $I_\Omega(a)$ , i.e.

$$a_1 := \arg \max_a I_\Omega(a).$$

Le premier axe principal est donc l'axe sur lequel la projection des données va avoir la dispersion la plus importante par rapport à l'individu de référence. D'une certaine manière, la direction associée à cet axe sera celle contenant le plus "d'informations". La proposition ci-dessous décrit les propriétés de ce premier axe principal.

#### Proposition 4.3.

Soit  $\Gamma$  la matrice introduite en (4.4). Le premier axe principal est un vecteur propre  $M$ -normé de  $\Gamma M$ . En particulier, il est solution de l'équation

$$(\Gamma M)a = \lambda a.$$

|| La valeur propre associée  $\lambda_1$  est la plus grande valeur propre de la matrice  $\Gamma M$  et vérifie

$$I_\Omega(a_1) = \lambda_1.$$

PREUVE :

On cherche à maximiser  $I_\Omega(a)$  sous la contrainte  $\|a\|_M^2 = a'Ma = 1$ . Exprimé sous forme lagrangienne, ce problème revient à chercher  $a_1, \rho_1$  tels que

$$(a_1, \rho_1) = \arg \max_{a, \rho} L(a, \rho) \quad \text{avec} \quad L(a, \rho) = I_\Omega(a) - \rho(a'Ma - 1).$$

La fonction  $L$  étant quadratique, l'unique maximum de  $L$  est obtenu pour  $(a_1, \rho_1)$  vérifiant

$$\begin{cases} \frac{\partial L}{\partial \rho}(a_1, \rho_1) = 0, \\ \frac{\partial L}{\partial a}(a_1, \rho_1) = 0. \end{cases} \quad (4.5)$$

On peut dans un premier temps remarquer que

$$I_\Omega(a) = \|TMA\|_W^2 = a'M'T'WTMA = a'M'\Gamma Ma.$$

En se rappelant que pour toute matrice symétrique  $G$  et vecteur  $h$ , on a

$$D_h(h'Gh) = 2Gh,$$

le système d'équations (4.5) peut alors être écrit sous la forme

$$\begin{cases} a'Ma - 1 = 0, \\ 2(M'\Gamma Ma - \rho Ma) = 0 \end{cases}$$

Dans la mesure où  $M$  est inversible, le vecteur  $a_1$  vérifie donc

$$M\Gamma Ma_1 - \rho_1 Ma_1 = 0 \quad \Leftrightarrow \quad \Gamma Ma_1 = \rho_1 a_1,$$

ce qui nous donne le premier élément de la proposition. Pour conclure la preuve, en posant  $\rho_1 = \lambda_1$ , il reste à remarquer que

$$I_\Omega(a_1) = a_1'M'\Gamma Ma_1 = \lambda_1 a_1'M'a_1 = \lambda_1,$$

dans la mesure où  $a_1$  est un vecteur  $M$ -normé. □

L'axe  $a_1$  associé à la plus grande valeur propre de la matrice  $\Gamma M$  est donc celui pour lequel la dispersion de la projection des données sur la droite associée est la plus importante. Une fois que cet axe a été identifié, il est possible d'effectuer la même manipulation sur les axes orthogonaux restants.

### 4.3.2 Décomposition en composantes principales

On supposera par la suite que les valeurs propres de la matrice  $\Gamma M$  sont toutes distinctes et rangées dans l'ordre décroissant :  $\lambda_1 > \dots > \lambda_p$ . Notons pour tout  $1 \leq j \leq p$ ,  $a_j$  le vecteur propre associé à la valeur propre  $\lambda_j$ .

Dans cette configuration, la proposition ci-dessous montre que les vecteurs propres associés sont tous deux-à-deux orthogonaux. Par ailleurs, l'inertie sur chacun des axes engendrés par ces vecteurs propres est égale à la valeur propre associée.

**Proposition 4.4.**

Soient  $\lambda_1 > \dots > \lambda_p$  les valeurs propres (supposées deux-à-deux distinctes) de  $\Gamma M$ . On désigne par  $a_1, \dots, a_p$  les vecteurs propres associés.

1. Pour tous  $j \neq k$ , on a

$$\langle a_j, a_k \rangle_M = 0.$$

2. De plus, pour tout  $j$  dans  $\{1, \dots, p\}$ , on a

$$I_\Omega(a_j) = \lambda_j.$$

PREUVE : Soient  $j \neq k$

① Par définition de  $(a_j)_j$ ,

$$\begin{cases} \Gamma M a_j = \lambda_j a_j \\ \Gamma M a_k = \lambda_k a_k \end{cases}$$

alors,  $\begin{cases} \langle \Gamma M a_j, a_k \rangle_M = \lambda_j \langle a_j, a_k \rangle_M \\ \langle a_j, \Gamma M a_k \rangle_M = \lambda_k \langle a_j, a_k \rangle_M \end{cases}$

car  $\begin{cases} M \text{ diagonale} \Rightarrow M' = M \\ \Gamma' = (T'WT)^T = T'WT = \Gamma \\ W = W' \end{cases}$

or  $\langle \Gamma M a_j, a_k \rangle_M = (\overset{\circ}{a_j} M \Gamma) \underset{\circ}{M a_k} = \langle a_j, \Gamma M a_k \rangle_M$

d'où  $\lambda_j \langle a_j, a_k \rangle_M = \lambda_k \langle a_j, a_k \rangle_M$  i.e.  $\underbrace{(\lambda_j - \lambda_k)}_{\neq 0} \underbrace{\langle a_j, a_k \rangle_M}_{=0} = 0$  d'où  $\langle a_j, a_k \rangle_M = 0$ .

②  $I_\Omega(a_j) = \|T M a_j\|_W^2 = \overset{\circ}{a_j} M \underbrace{T' W T M a_j}_M = \overset{\circ}{a_j} M \Gamma M a_j$

$$= \overset{\circ}{a_j} M \Gamma M a_j$$

$$= \overset{\circ}{a_j} M \lambda_j a_j = \lambda_j \underbrace{\|a_j\|_M^2}_{=1} = \lambda_j .$$

car  $a_j$  est  $M$ -orthogonal

□

Il est possible de montrer que  $a_2$  est l'axe maximisant l'inertie parmi tous les axes orthogonaux à  $a_1$ , et ainsi de suite. La décomposition en composantes principales revient donc à exhiber les valeurs et vecteurs propres  $(\lambda_j, a_j)_{j=1 \dots p}$  de la matrice  $\Gamma M$ , ce qui revient d'une certaine manière à effectuer un changement de base.

**Définition 4.4.**

Soient  $(\lambda_j, a_j)_{j=1 \dots p}$  les valeurs propres et vecteurs propres de la matrice  $\Gamma M$  :

- Les vecteurs propres  $a_j \in \mathbb{R}^p$  sont appelés **axes (factoriels) principaux**.
- Les vecteurs  $u_j := Ma_j \in \mathbb{R}^p$  sont appelés les **facteurs principaux**.
- Les vecteurs  $c_j := Tu_j$  sont appelées les **composantes principales**.
- Pour  $k, j \in \{1, \dots, p\}$ ,  $k \neq j$ , le plan engendré par les vecteurs  $a_j$  et  $a_k$  est appelé **plan factoriel** ( $a_j, a_k$ ).

Remarquons que la *je* métavariable vaut

$$c_j = Tu_j = TMa_j = \begin{pmatrix} \langle T_1, a_j \rangle_M \\ \vdots \\ \langle T_n, a_j \rangle_M \end{pmatrix} \in \mathbb{R}^n.$$

En effet,

$$\begin{aligned} c_j &= TMa_j = \begin{pmatrix} T_1 \\ \vdots \\ T_n \end{pmatrix} \xrightarrow{\text{proj}} (Ma_j) \xrightarrow{\text{proj}} \\ &= \begin{pmatrix} T_1 Ma_j \\ \vdots \\ T_n Ma_j \end{pmatrix} = \begin{pmatrix} \langle T_1, a_j \rangle_M \\ \vdots \\ \langle T_n, a_j \rangle_M \end{pmatrix} \xrightarrow{\text{proj}} \end{aligned}$$

En particulier, par la Proposition 4.4,

$$\|c_j\|_W^2 = \|TMA_j\|_W^2 = I_\Omega(a_j) = \lambda_j.$$

D'une certaine manière, les axes (facteurs) principaux définissent un nouveau repère dans  $\mathbb{R}^p$ , en lieu et place de celui engendré par la base canonique. De même, les projections du nuage de points sur ces nouveaux axes (i.e. les composantes principales), définissent ce qu'il conviendrait d'appeler des **meta-variables**. Autrement dit, nous sommes partis d'un tableau de données où chaque colonne représentait les valeurs d'une variable en question (avec une inertie associée) pour arriver à une nouvelle représentation permettant une meilleure "explication" de l'inertie.

**En pratique**

En reprenant la matrice  $T$  définie en (4.2), on rappelle que

$$\Gamma = \Gamma M = \begin{pmatrix} 11.39 & 9.92 & 2.66 & 4.82 \\ 9.92 & 8.94 & 4.12 & 5.48 \\ 2.66 & 4.12 & 12.06 & 9.29 \\ 4.82 & 5.48 & 9.29 & 7.91 \end{pmatrix}.$$

A l'aide de n'importe quel logiciel mathématique (e.g. logiciel R, commande `eigen`), il est possible de récupérer les valeurs propres et vecteurs propres de cette matrice  $\Gamma$ . On obtient ici

$$\lambda_1 = 28.23, \quad \lambda_2 = 12.03, \quad \lambda_3 = 0.03, \quad \lambda_4 = 0.01.$$

Les axes (facteurs) principaux associés sont ici

$$a_1 = \begin{pmatrix} 0.52 \\ 0.51 \\ 0.49 \\ 0.48 \end{pmatrix}, \quad a_2 = \begin{pmatrix} -0.57 \\ -0.37 \\ 0.66 \\ 0.33 \end{pmatrix}, \quad a_3 = \begin{pmatrix} 0.19 \\ -0.45 \\ -0.46 \\ 0.74 \end{pmatrix}, \quad a_4 = \begin{pmatrix} -0.61 \\ 0.63 \\ -0.34 \\ 0.33 \end{pmatrix}.$$

Il est ensuite possible de calculer les 4 composantes principales  $c_j$  en appliquant la formule  $c_j = Ta_j$  (car  $u_j = Ma_j = a_j$ ) :

	$c_1$	$c_2$	$c_3$	$c_4$
Nathan	-8.61	-1.41	0.07	0.07
Emma	-3.88	-0.50	0.01	-0.07
Lola	-3.21	3.47	-0.17	0.01
Baptiste	9.85	0.60	0.04	-0.15
Mathilde	6.41	-2.05	-0.08	0.19
Ines	-3.03	-4.92	0.08	-0.14
Lucas	-1.03	6.38	-0.16	-0.03
Aymeric	1.95	-4.20	-0.20	0.04
Chloe	1.55	2.63	0.42	0.07

Chaque colonne  $c_j$  représente une métavariable.

Dans le contexte qui nous intéresse, les  $c_j$  peuvent être vues comme des **méta-matières** et les entrées correspondantes comme des scores (méta-notes) dans chacune de ces métamatières (auxquelles il va maintenant falloir donner une signification).

## 4.4 Interprétation des résultats

Le but principal de l'ACP étant de présenter une description synthétique du jeu de données considéré, il n'est pas pertinent de travailler avec l'ensemble des métavariables obtenues par la méthode précédente. Il va donc falloir dans un premier temps essayer de sélectionner un "petit" nombre de variables explicatives. Dans un second temps, il conviendra de donner une signification à ces métavariables.

### 4.4.1 Choix du nombre de composantes

Comme expliqué plus haut, le but de l'ACP est de proposer une représentation des données permettant d'expliquer la dispersion du nuage de points (l'inertie) de la manière la plus synthétique possible. A ce titre, les valeurs propres de la matrice  $\Gamma M$  apportent beaucoup d'informations.

#### Proposition 4.5.

Soient  $\lambda_1 > \dots > \lambda_p$  les valeurs propres de la matrice  $\Gamma M$ . Alors

$$I_\Omega = \sum_{j=1}^p \lambda_j.$$

PREUVE : alternative :  $(a_j)_{1 \leq j \leq p}$  est une base  $M - \mathbb{I} \mathbb{I} \mathbb{I}$  de  $\mathbb{R}^p$

D'après la Proposition 4.1,

$$I_\Omega = \text{Tr}(\Gamma M).$$

$$\Rightarrow I_\Omega = \sum_{j=1}^p \overbrace{\text{In}_\Omega(a_j)}^{\text{prop 4.4}} = \sum_{j=1}^p \lambda_j. \quad \text{prop 4.2}$$

La trace d'une matrice étant invariante par changement de base, on obtient le résultat proposé ci-dessus.  $\square$

On sait depuis la Proposition 4.4 que chaque axe  $a_j$  "porte" une inertie  $\lambda_j$ . Par ailleurs, grâce à la Proposition 4.5, l'inertie totale du nuage de points est égale à la somme des inerties portées par chacun des axes  $a_j$  (propriété valable quelque soit le repère).

Dans ce contexte, il semble "naturel" de ne conserver que les axes associés à une part significative de l'inertie. Plus formellement, on définit pour tout  $j$  dans  $\{1, \dots, p\}$ ,

$$PC(j) = \sum_{l=1}^j \frac{\lambda_l}{I_\Omega} = \frac{\sum_{l=1}^j \lambda_l}{\sum_{l=1}^p \lambda_l}.$$

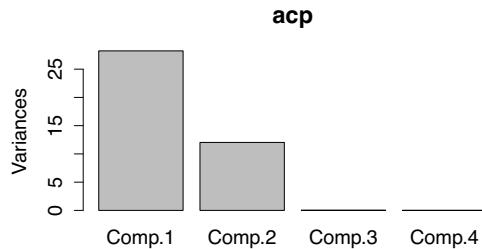
Pour tout entier  $j$  compris entre 1 et  $p$ , l'indice  $PC(j)$  désigne le pourcentage d'inertie cumulé sur les  $j$  premiers axes principaux. Dans la plupart des applications courantes, le nombre  $j_0$  de composantes principales retenues sera le plus petit  $j$  tel que

$$PC(j) \geq 1 - \alpha,$$

où typiquement  $\alpha$  est compris entre 10% et 20%. A noter que d'autres types de critères peuvent être sélectionnés comme par exemple d'une rupture "nette" dans la décroissance des valeurs propres.

### En pratique

Le graphique suivant rassemble l'inertie associée à chacune des composantes (i.e. les valeurs propres)



On se rend vite compte que l'essentiel de l'inertie est concentrée sur les deux premiers axes principaux. Si on regarde la représentation des différentes composantes en terme de pourcentage d'inertie, on obtient le tableau suivant

Composante	Pourcentage d'inertie	Pourcentage cumulé d'inertie
$c_1$	70.05%	70.05 %
$c_2$	29.84%	99.89 %
$c_3$	0.08%	99.97 %
$c_4$	0.03%	100 %

Ainsi, les deux premières composantes principales cumulent à elles seules plus de 99% de l'inertie initiale. Dans ce contexte, les deux dernières composantes apparaissent comme peu informatives et peuvent donc être abandonnées sans regret. Il reste maintenant à donner une signification à ces deux premières composantes principales...

Enfin, il convient d'essayer de donner une signification aux composantes retenues pour la description des données. On va utiliser pour cela plusieurs outils, en s'intéressant d'une part aux individus les plus "représentatifs" d'une composante principale donnée, et en mesurant d'autre part la corrélation entre les variables initiales et les métavariables obtenues après diagonalisation de la matrice.

#### 4.4.2 Liens entre les individus et les axes principaux

##### a) Graphe des projections orthogonales des individus

Il arrive parfois que certains individus aient des "profils" très particuliers, permettant ainsi d'affiner l'interprétation des différentes composantes principales. Plus formellement, une composante principale donnée  $c_k$  correspond à la projection  $M$ -orthogonale du nuage de points initial sur l'axe principal associé  $a_k$ . En particulier, la  $i^{\text{ème}}$  coordonnée de  $c_k$  est donnée par

$$(c_k)_i = \langle T_i, a_k \rangle_M.$$

Graphiquement, on représente les projections  $M$ -orthogonales du nuage de points (dans  $\mathbb{R}^p$ ) sur le premier plan factoriel (engendré par  $a_1$  et  $a_2$ ). Plus précisément, chaque individu  $i$  est représenté par les points de coordonnées

$$\left( \langle T_i, a_1 \rangle_M, \langle T_i, a_2 \rangle_M \right) = \left( (c_1)_i, (c_2)_i \right),$$

pour former le *graphe des individus*.

Il est parfois utile également de représenter des projections sur d'autres plans factoriels (engendrés par  $a_k$  et  $a_l$ , pour  $1 \leq k \neq l \leq p$ ).

##### b) Contribution relative des individus

Rappelons que l'inertie portée par l'axe  $a_k$  est définie comme la somme des distances entre les points projetés sur cet axe et l'individu de référence. En particulier, on a

$$\lambda_k = I_\Omega(a_k) = \sum_{i=1}^n w_i \langle T_i, a_k \rangle_M^2.$$

A ce titre, les individus les plus intéressants seront ceux qui sont les plus éloignés de cet individu moyen, i.e. ceux dont la "contribution" à la dispersion est la plus importante. Cette heuristique motive la définition suivante.

##### Définition 4.5.

La contribution relative de l'individu  $i$  à l'inertie de  $c_k$  (i.e. sur l'axe  $a_k$ ) est définie par

$$\text{CTR}_{\text{ind}}(i, k) := \frac{w_i \langle T_i, a_k \rangle_M^2}{\lambda_k} = \frac{w_i (c_k)_i^2}{\lambda_k}.$$

On a bien  $\sum_{i=1}^n \text{CTR}_{\text{ind}}(i, k) = 1$ .

En pratique, il conviendra de s'intéresser au "profil" de individus ayant la contribution relative la plus importante afin d'essayer d'interpréter la signification d'une composante principale donnée.

##### c) Qualité de représentation des individus

La qualité de représentation d'un individu  $i$  sur l'axe  $k$  peut être mesurée par la distance entre le point dans l'espace et sa projection sur l'axe. On préfère en réalité représenter le pourcentage d'inertie de l'individu  $i$  projeté sur l'axe engendré par  $a_k$  (par rapport à son inertie initiale).

##### Définition 4.6.

La qualité de représentation de l'individu  $i$  sur l'axe  $k$  est définie par

$$qlt(i, k) = \frac{\langle T_i, a_k \rangle_M^2}{\|T_i\|_M^2}.$$

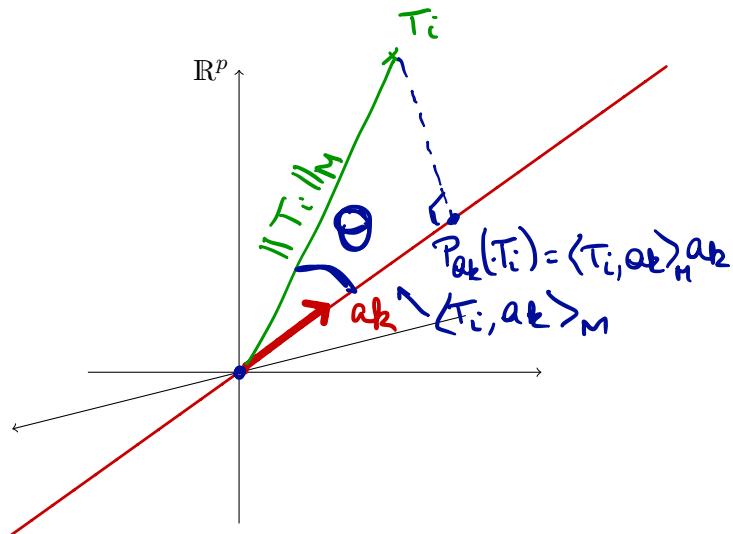
Remarquons que, d'un point de vue géométrique, la qualité de représentation n'est autre que le carré du cosinus de l'angle entre le vecteur  $T_i$  et l'axe  $a_k$ .

En effet,

$$\cos(\theta) = \frac{\langle T_i, a_k \rangle_M}{\|T_i\|_M}.$$

d'où

$$\text{qlt}(i, k) = \cos^2(\theta)$$



Pour cette raison, la qualité de représentation est parfois appelée le cosinus carré, ce qui est le cas dans le package FactoMineR que nous utiliserons en TP.

#### 4.4.3 Liens entre les variables et les composantes principales

##### a) Corrélation entre les variables initiales et les composantes principales

Les variables initiales sont des éléments de  $\mathbb{R}^n$ , tout comme les composantes principales. La proposition suivante met en avant un lien entre ces deux types de quantités.

##### Proposition 4.6.

*Les composantes principales sont des combinaisons linéaires des variables initiales. Plus formellement, pour tout  $j \in \{1, \dots, p\}$ ,*

$$c_j = \sum_{k=1}^p u_{j,k} T^{(k)},$$

où  $u_{j,k}$  désigne la  $k^{\text{ème}}$  coordonnée du facteur  $u_j$ .

En particulier, si le tableau  $T$  est centré, les composantes principales sont également centrées.

PREUVE :

Rappelons que

$$g_j = TM_{\text{maj}} = Tu_j = \left( \begin{array}{c} \downarrow \\ n \end{array} \right) \left( \begin{array}{c} \leftarrow \\ T^{(1)} \dots T^{(p)} \end{array} \right) \left( \begin{array}{c} \uparrow \\ p \end{array} \right) = \left( \begin{array}{c} u_{j,1} \\ \vdots \\ u_{j,p} \end{array} \right) \left( \begin{array}{c} \uparrow \\ p \end{array} \right)$$

$$= u_{j,1} T^{(1)} + \dots + u_{j,p} T^{(p)} \in \mathbb{R}^n = \sum_{k=1}^p u_{j,k} T^{(k)} \quad ( \in \mathbb{R}^n ).$$

□

Dans la mesure où les composantes principales sont des combinaisons linéaires des variables initiales, il apparaît pertinent de s'intéresser à leurs coefficients de corrélation.

Remarquons que, dans le cas standard où  $W = n^{-1}I_n$ , pour toutes variables *centrées*  $x$  et  $y$  de  $\mathbb{R}^n$ ,

$$\text{Var}(x) = \|x\|_W^2 \quad \text{et} \quad \text{Cov}(x, y) = \langle x, y \rangle_W.$$

En effet,

$$\begin{aligned} \langle x, y \rangle_W &= \sum_{i=1}^n \frac{x_i}{y_n} x_i y_i = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \text{cov}(x, y) \\ \|x\|_W^2 &= \langle x, x \rangle_W = \text{cov}(x, x) = \text{var}(x). \end{aligned}$$

En particulier, nous obtenons immédiatement la propriété suivante.

#### Propriété 4.1.

Si  $T$  est un tableau centré, alors le coefficient de corrélation entre la variable initiale  $T^{(j)}$  et la composante principale  $c_k$  vaut

$$r_{j,k} = \frac{\text{Cov}(T^{(j)}, c_k)}{\sqrt{\text{Var}(T^{(j)}) \text{Var}(c_k)}} = \frac{\langle T^{(j)}, c_k \rangle_W}{\|T^{(j)}\|_W \|c_k\|_W}.$$

Graphiquement, on représente chaque variables  $T^{(j)}$  dans le cercle des corrélations par un point de coordonnées

$$\left( r_{j,1}, r_{j,2} \right) = \left( \frac{\langle T^{(j)}, c_1 \rangle_W}{\|T^{(j)}\|_W \|c_1\|_W}, \frac{\langle T^{(j)}, c_2 \rangle_W}{\|T^{(j)}\|_W \|c_2\|_W} \right),$$

pour former le *graphe des variables*.

En particulier, si une variable est proche d'un axe, elle est fortement corrélée positivement ou négativement avec la méta-variable correspondante. Ainsi, pour donner une signification à une composante, on regardera surtout les variables qui sont alignées avec l'axe correspondant, i.e. pour lesquelles le coefficient de corrélation sera "proche" de 1 en valeur absolue.

Remarquons enfin qu'une corrélation entre deux variables peut être interprétée comme un cosinus, FactoMineR calcule également les cosinus carrés entre les variables et les composantes comme étant le carré des corrélations correspondantes.

#### b) Contribution d'une variable à la construction d'une composante principale

Par ailleurs, il semble également pertinent de chercher à savoir quelles sont les variables initiales qui apportent les contributions les plus importantes dans les valeurs d'une composante principale donnée. Remarquons que la variance de la composante principale  $c_k$  se décompose à l'aide de ses covariances avec les variables initiales de la manière suivante.

#### Lemme 4.1.

Si le tableau  $T$  est centré, alors pour tout  $1 \leq k \leq p$ ,

$$\sum_{j=1}^p m_j \langle T^{(j)}, c_k \rangle_W^2 = \lambda_k^2.$$

PREUVE : Calculons d'abord

$$\begin{aligned} \langle T^{(j)}, c_k \rangle_W &= T^{(j)'} W c_k = T^{(j)'} W \underbrace{T M a_k}_{\Gamma} = \underbrace{(T' W T M a_k)}_{\Gamma} \\ &= (\Gamma M a_k)_j = \lambda_k a_{k,j} \end{aligned}$$

$$\text{d'où } \sum_{j=1}^p m_j \langle T^{(j)}, c_k \rangle_W^2 = \lambda_k^2 \sum_{j=1}^p m_j a_{k,j}^2 = \lambda_k^2 \underbrace{\|a_k\|_M^2}_{=1} = \lambda_k^2.$$

□

Comme dans le cas des individus, ceci motive la définition suivante.

#### Définition 4.7.

Dans le cas d'un tableau centré, la contribution de la variable  $T^{(j)}$  à la construction de la composante principale  $c_k$  est définie par

$$\text{CTR}_{var}(j, k) = \frac{m_j \langle T^{(j)}, c_k \rangle_W^2}{\lambda_k^2} = \frac{m_j \|T^{(j)}\|_W^2 r_{j,k}^2}{\lambda_k}.$$

$$\text{En effet, } r_{j,k}^2 = \frac{\langle T^{(j)}, c_k \rangle_W^2}{\|T^{(j)}\|_W^2 \|c_k\|_W^2} \quad \text{avec } \|c_k\|_W^2 = \|T M a_k\|_W^2 = I_{\text{Ind}}(a_k) = \lambda_k$$

$$\text{d'où } \langle T^{(j)}, c_k \rangle_W^2 = \|T^{(j)}\|_W^2 \lambda_k r_{j,k}^2$$

Comme dans le cas des individus, on a bien  $\sum_{j=1}^p \text{CTR}_{var}(j, k) = 1$ .

Remarquons par ailleurs que l'on peut faire le rapprochement avec le cas des individus grâce au cosinus carré.

En effet, cas de variable :

$$\cos^2(j, k) = r_{j,k}^2$$

$$\Rightarrow \text{CTR}_{var}(j, k) = \frac{m_j \|T^{(j)}\|_W^2 \cos^2(j, k)}{\lambda_k}$$

$$\left. \begin{array}{l} \text{cas des individus :} \\ \cos^2(j, k) = q_{lt}(j, k) = \frac{\langle T_i, a_k \rangle_n^2}{\|T_i\|_n^2} \\ \Rightarrow \text{CTR}_{ind}(i, k) = \frac{\omega_i \|T_i\|_n^2 \cos^2(i, k)}{\lambda_k} \end{array} \right\}$$

**Cas particulier centré et réduit** Dans le cas particulier d'un tableau  $T$  centré réduit, avec  $M = I_p$ , on a pour tous  $j$  que  $\|T^{(j)}\|_W = 1$  d'où la contribution s'écrit

$$\text{CTR}_{var}(j, k) = \frac{r_{j,k}^2}{\lambda_k}.$$

Remarquons que c'est également le cas lorsque le tableau  $T$  est centré et que la pondération par les variances est dans la matrice  $M$ .

Ainsi, dans ce cas, étudier la contribution revient à étudier la corrélation. Il devient alors intéressant d'évaluer la contribution des variables surtout en ACP centrée non réduite.

### En pratique

Nous avons maintenant tous les outils nécessaires à la mise en place d'une analyse en composantes principales du tableau 4.1. On rappelle que suite aux discussions précédentes, seules deux composantes principales ont été retenues dans l'analyse.

	$c_1$	$c_2$
Nathan	-8.61	-1.41
Emma	-3.88	-0.50
Lola	-3.21	3.47
Baptiste	9.85	0.60
Mathilde	6.41	-2.05
Ines	-3.03	-4.92
Lucas	-1.03	6.38
Aymeric	1.95	-4.20
Chloe	1.55	2.63

Il reste maintenant à donner une signification à  $c_1$  et  $c_2$ . On peut remarquer à l'aide de la table précédente que Nathan et Baptiste ont un "profil" très marqué sur la première composante, alors que sur la seconde, Ines et Lucas sont les deux étudiants les plus discriminants.

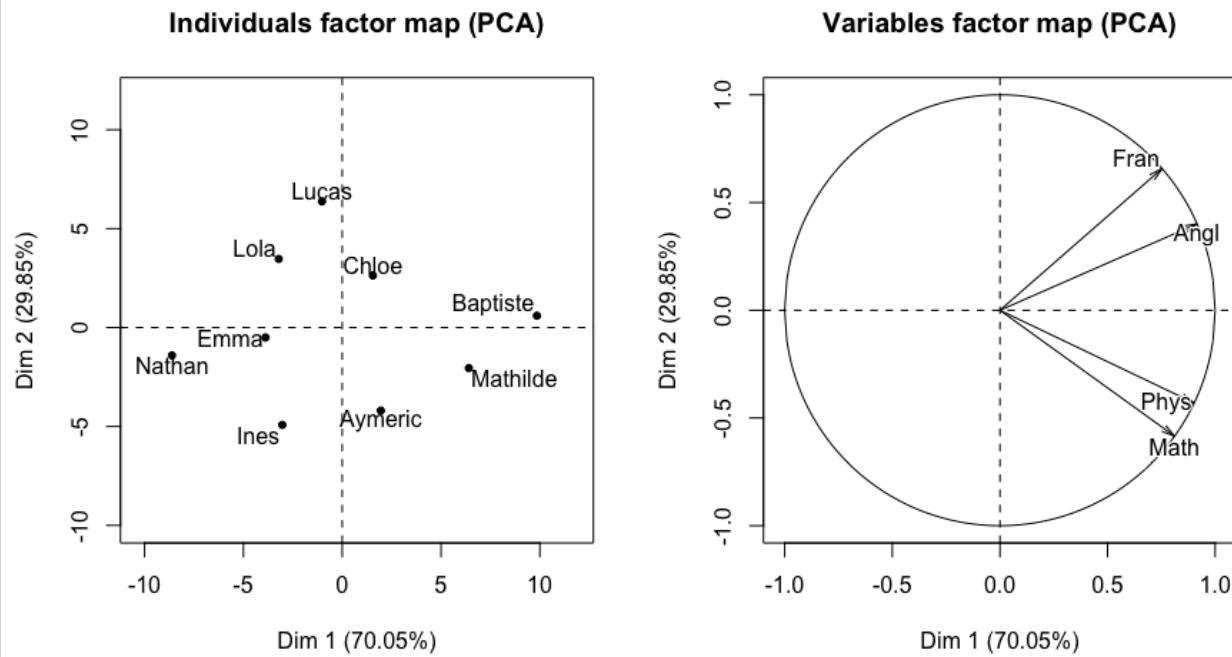
Un rapide coup d'oeil à la table 4.1 peut permettre de se faire une première idée : Baptiste est l'étudiant ayant les meilleurs résultats alors que Nathan est plus en difficulté, Ines a des notes correctes en Mathématiques et Physique mais c'est moins le cas en Français et Anglais... tout au contraire de Lucas.

Afin d'aller plus loin dans l'analyse, on peut maintenant s'intéresser à la corrélation entre les deux composantes  $c_1$ ,  $c_2$  et les variables initiales. Ces coefficients sont regroupés dans le tableau suivant :

	$c_1$	$c_2$
Math	0.81	-0.58
Physique	0.90	-0.43
Français	0.75	0.66
Anglais	0.91	0.40

La composante  $c_1$  est corrélée négativement avec l'ensemble des variables de départ : un individu ayant de bonnes notes dans les quatre matières se verra typiquement assigné un faible coefficient dans la première composante. La composante 2 est corrélée positivement aux matières "scientifiques", mais négativement aux matières "littéraires".

Dans la mesure où on ne retient que deux composantes pour expliquer le nuage de points, il est possible de synthétiser la discussion précédente à l'aide du graphique suivant (obtenus grâce au package FactoMineR sous .



Dans le graphe des individus (à gauche), on représente les projections des différents individus sur le premier plan factoriel (i.e. le plan engendré par les deux premiers axes principaux  $a_1$  et  $a_2$ ). Nous pouvons vérifier que les coordonnées sont contenues dans les deux premières composantes principales.

Dans le graphe des variables, on représente les corrélations entre les variables initiales et les deux premières composantes principales.

**Bilan :** On peut résumer le nuage de points à l'aide de deux métavariables. La première composante mesure le niveau global de l'étudiant alors que la seconde mesure le profil scientifique/littéraire.





