UNIVERSITE
PAUL
SABATIER
TOULOUSE III

Publications du Laboratoire
de
Statistique et Probabilités

# Statistique Descriptive

Alain BACCINI

Laboratoire de Statistique et Probabilités — UMR CNRS C5583
Université Paul Sabatier — 31062 – Toulouse cedex 4.

# Sommaire

# Chapitre 1

# Introduction

Ce cours de statistique descriptive, ou exploratoire, est organisé en deux documents.

- Le premier et présent document propose des éléments de statistique descriptive unidimensionnelle et bidimensionnelle ; il ne nécessite aucun outil mathématique spécifique et peut être enseigné en premier cycle universitaire.
- Le deuxième document (Baccini et Besse) fait largement appel au calcul matriciel dont les éléments fondamentaux sont rappelés en annexe A ; il expose les fondements des techniques factorielles mises en œuvre dans la plupart des logiciels statistiques et correspond à un enseignement de second cycle.

Les graphiques présentés ont été réalisés à l'aide des logiciels SAS ou S+. Les programmes informatiques mentionnés, ainsi que les mises à jours des documents cités, sont accessibles sur le site internet suivant :

`http://www-sv.cict.fr/lsp/Besse/`

## 1 Généralités sur la statistique

### Définition

Il n'est pas commode, dans cette introduction, de donner une définition précise du concept de statistique, alors que son contenu sera en partie élaboré dans la suite de ce cours. Nous nous contenterons donc, pour fixer les idées, d'en donner une définition volontairement assez vague.

**Définition 1.1** *On appelle Statistique l'ensemble des méthodes (ou encore des techniques) permettant d'analyser (on dira plutôt de* traiter*) des ensembles d'observations (nous parlerons de* données*).*

Les méthodes en question relèvent le plus souvent des mathématiques (raison pour laquelle, au moins en France, la Statistique fait partie des mathématiques appliquées) et font largement appel à l'outil informatique pour leur mise en œuvre.

On notera la distinction entre *la* Statistique, au sens défini ci-dessus, et *une* statistique, terme parfois utilisé pour désigner des "données statistiques" (voir ce terme plus loin) ; par exemple, on parle de la statistique du commerce extérieur de la France. Dans la suite de ce cours, nous n'utiliserons pas le terme de statistique dans ce dernier sens.

**Bref historique**

De façon un peu grossière, on peut distinguer trois phases essentielles dans l'évolution de la statistique.

- Depuis l'antiquité et jusqu'à la fin du $19^{\text{ième}}$ siècle, la statistique est restée principalement un ensemble de techniques de dénombrement.
- Entre la fin du $19^{\text{ième}}$ siècle et les années 1960, s'est construit, notamment à la suite de l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman...), la statistique mathématique (ou statistique inférentielle, voir ci-dessous).
- Depuis les années 1960, avec le développement des outils informatiques et graphiques, la statistique, et surtout la statistique descriptive multidimensionnelle, a connu une expansion considérable.

**Statistique descriptive et statistique inférentielle**

De manière approximative, il est possible de classer les méthodes statistiques en deux groupes : celui des méthodes descriptives et celui des méthodes inférentielles.

- La statistique **descriptive**. On regroupe sous ce terme les méthodes dont l'objectif principal est la *description* des données étudiées ; cette description des données se fait à travers leur *présentation* (la plus synthétique possible), leur *représentation graphique*, et le calcul de *résumés numériques*. Dans cette optique, il n'est pas fait appel à des modèles probabilistes. On notera que les termes de statistique descriptive, *statistique exploratoire* et *analyse des données* sont quasiment synonymes. C'est essentiellement à ces méthodes qu'est consacré ce cours.
- La statistique **inférentielle**. Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population ; d'une certaine manière, il s'agit donc d'induire (ou encore d'inférer) du particulier au général. Le plus souvent, ce passage ne pourra se faire que moyennant des hypothèses de type probabiliste. Les termes de statistique inférentielle, *statistique mathématique*, et *statistique inductive* sont eux aussi quasiment synonymes.

D'un point de vue méthodologique, on notera que la statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données : les deux aspects de la statistique se complètent bien plus qu'ils ne s'opposent.

## 2 Terminologie de base

**Population** $\Omega$ (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de *champ de l'étude*.

**Individu** $\omega \in \Omega$ (ou *unité statistique*) : tout élément de la population.

**Échantillon** : sous–ensemble de la population sur lequel sont effectivement réalisées les observations.

**Taille de l'échantillon** $n$ : cardinal du sous-ensemble correspondant.

**Enquête** (statistique) : opération consistant à observer (ou mesurer, ou questionner...) l'ensemble des individus d'un échantillon.

**Recensement** : enquête dans laquelle l'échantillon observé est la population tout entière (enquête *exhaustive*).

**Sondage** : enquête dans laquelle l'échantillon observé est un sous–ensemble strict de la population (enquête *non exhaustive*).

**Variable** (statistique) : $\Omega \stackrel{X}{\longmapsto} \begin{cases} \mathcal{E} \text{ si qualitative} \\ \mathbb{R} \text{ si quantitative} \end{cases}$
caractéristique (âge, salaire, sexe...), définie sur la population et observée sur l'échantillon ; mathématiquement, il s'agit d'une application définie sur l'échantillon. Si la variable est à valeurs dans $\mathbb{R}$ (ou une partie de $\mathbb{R}$, ou un ensemble de parties de $\mathbb{R}$), elle est dite *quantitative* (âge, salaire, taille...) ; sinon elle est dite *qualitative* (sexe, catégorie socioprofessionnelle...).

**Données** (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Elles sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Lorsqu'un tableau ne comporte que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de *matrice*.

# Chapitre 2

# Cas unidimensionnel

Si $X$ est une variable statistique et si $\omega_i$ désigne l'individu générique de l'échantillon observé, nous noterons $X(\omega_i)$ la valeur prise par cette variable sur cet individu. C'est dorénavant l'échantillon observé, qu'il soit identique à la population complète ou non, qui sera noté $\Omega$ ; nous le supposerons de cardinal $n$. L'ensemble $\{X(\omega_i)\ ;\ i = 1, \ldots, n\}$ constitue ce que l'on appelle la *série statistique brute* ; c'est en général sous cette forme que se présentent les données dans un fichier informatique. Elles sont alors parfaitement illisibles dès que $n$ est grand ; l'objectif de ce chapitre est d'exposer les outils élémentaires, adaptés au type de variable observée, permettant de présenter une série brute de façon synthétique et d'en résumer les principales caractéristiques. Sont ainsi introduites les notions de médiane, quantile, moyenne, variance, écart–type parallèlement aux représentations graphiques usuelles : diagramme en bâton, histogramme, boîte–à–moustaches, graphiques cumulatifs, diagrammes en colonnes, en barre ou en secteurs.

## 1 Variable quantitative discrète

### 1.1 Introduction

En général, on appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs entières (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible (sauf exception, moins d'une vingtaine). Citons, par exemple, le nombre d'enfants dans une population de familles, le nombre d'années d'études après le bac dans une population d'étudiants...

**Exemple 2.1** *On a noté l'âge (arrondi à l'année près) des 48 salariés d'une entreprise ; la série statistique brute est donnée ci-dessous (il s'agit de données fictives).*

43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31
62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59
57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

### 1.2 Présentation des données

#### Le tableau statistique

C'est un tableau dont la première colonne comporte l'ensemble des $r$ observations distinctes de la variable $X$ ; ces observations sont rangées par ordre croissant et non répétées ; nous les noterons $\{x_l\ ;\ l = 1, \ldots, r\}$. Dans une seconde colonne, on dispose, en face de chaque valeur $x_l$, le nombre de

| $x_l$ | $n_l$ | $N_l$ | $f_l(\%)$ | $F_l(\%)$ |
|---|---|---|---|---|
| 24 | 1 | 1 | 2,08 | 2,08 |
| 26 | 2 | 3 | 4,17 | 6,25 |
| 29 | 3 | 6 | 6,25 | 12,50 |
| 31 | 2 | 8 | 4,17 | 16,67 |
| 33 | 4 | 12 | 8,33 | 25,00 |
| 37 | 2 | 14 | 4,17 | 29,17 |
| 38 | 4 | 18 | 8,33 | 37,50 |
| 41 | 3 | 21 | 6,25 | 43,75 |
| 43 | 3 | 24 | 6,25 | 50,00 |
| 45 | 1 | 25 | 2,08 | 52,08 |
| 46 | 6 | 31 | 12,50 | 64,58 |
| 49 | 3 | 34 | 6,25 | 70,83 |
| 50 | 1 | 35 | 2,08 | 72,91 |
| 52 | 3 | 38 | 6,25 | 79,16 |
| 57 | 5 | 43 | 10,42 | 89,58 |
| 59 | 2 | 45 | 4,17 | 93,75 |
| 60 | 2 | 47 | 4,17 | 97,92 |
| 62 | 1 | 48 | 2,08 | 100,00 |

TAB. 2.1 – Effectifs, effectifs cumulés, fréquences et fréquences cumulées.

réplications qui lui sont associées ; ces réplications sont appelées *effectifs* et notées $n_l$. Les effectifs $n_l$ sont souvent remplacés par les quantités $f_l = \frac{n_l}{n}$ , appelées *fréquences* (rappelons que $n$ désigne le nombre total d'observations, c'est–à–dire le cardinal de $\Omega$ : $n = \sum_{l=1}^{r} n_l$).

**Les effectifs cumulés et les fréquences cumulées**

Il peut être utile de compléter le tableau statistique en y rajoutant soit les effectifs cumulés, soit les fréquences cumulées. Ces quantités sont respectivement définies de la façon suivante :

$$N_l = \sum_{j=1}^{l} n_j \text{ et } F_l = \sum_{j=1}^{l} f_j .$$

On notera que $N_r = n$ et $F_r = 1$.

**Illustration**

Dans le tableau statistique (2.1), on a calculé, sur les données présentées dans l'exemple 2.1, les effectifs, effectifs cumulés, fréquences et fréquences cumulées.

**Remarques 2.1**

- *Comme c'est le cas ci-dessus, les fréquences sont souvent exprimées en pourcentages.*
- *Le choix entre effectifs (resp. effectifs cumulés) et fréquences (resp. fréquences cumulées) est très empirique ; il semble naturel de choisir les effectifs lorsque l'effectif total $n$ est faible et les fréquences lorsqu'il est plus important ; la limite approximative de* 100 *paraît, dans ces conditions, assez raisonnable.*

**La présentation tige–et–feuille (ou "stem–and–leaf")**

Cette façon particulière de présenter les données est assez commode, dans la mesure où elle préfigure déjà un graphique. Elle est illustrée ci–dessous sur le même exemple que précédemment.

| | |
|---|---|
| 2 | 4 6 6 9 9 9 |
| 3 | 1 1 3 3 3 3 7 7 8 8 8 8 |
| 4 | 1 1 1 3 3 3 5 6 6 6 6 6 6 9 9 9 |
| 5 | 0 2 2 2 7 7 7 7 7 9 9 |
| 6 | 0 0 2 |

Elle consiste donc, dans la présentation des données, à séparer la partie des dizaines de celle des unités. En face de la partie des dizaines, chaque unité est répétée autant de fois qu'il y a d'observations de la valeur corresponsdante. Bien entendu, cette présentation doit être adaptée de façon appropriée lorsque les données sont d'un autre ordre de grandeur.

## 1.3 Représentations graphiques usuelles

Pour une variable discrète, on rencontre essentiellement deux sortes de représentations graphiques, qui sont en fait complémentaires: le diagramme en bâtons et le diagramme cumulatif (en escaliers).

**Le diagramme en bâtons**

Il permet de donner une vision d'ensemble des observations réalisées. La figure 2.1 donne le diagramme en bâtons des données de l'exemple 2.1.

**Le diagramme cumulatif**

Il figure les effectifs cumulés (resp. les fréquences cumulées) et permet de déterminer simplement le nombre (resp. la proportion) d'observations inférieures ou égales à une valeur donnée de la série. Lorsqu'il est relatif aux fréquences, c'est en fait le graphe de la *fonction de répartition empirique* $F_X$ définie de la façon suivante :

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1, \\ F_l & \text{si } x_l \leq x < x_{l+1}, \quad l = 1, \ldots, r-1, \\ 1 & \text{si } x \geq x_r. \end{cases}$$

Le diagramme cumulatif relatif à l'exemple 2.1 est donné par la figure 2.2.

## 1.4 Notion de quantile et applications

**Définition**

La fréquence cumulée $F_l$ $(0 \leq F_l \leq 1)$ donne la proportion d'observations inférieures ou égales à $x_l$. Une approche complémentaire consiste à se donner a priori une valeur $\alpha$, comprise entre 0 et 1, et à rechercher $x_\alpha$ vérifiant $F_X(x_\alpha) \simeq \alpha$. La valeur $x_\alpha$ (qui n'est pas nécessairement unique) est appelée quantile (ou *fractile*) d'ordre $\alpha$ de la série. Les quantiles les plus utilisés sont associés à certaines valeurs particulières de $\alpha$.

**La médiane et les quartiles**

La médiane est le quantile d'ordre $\frac{1}{2}$; elle partage donc la série des observations en deux ensembles d'effectifs égaux. Le premier quartile est le quantile d'ordre $\frac{1}{4}$, le troisième quartile celui d'ordre $\frac{3}{4}$ (le second quartile est donc confondu avec la médiane).

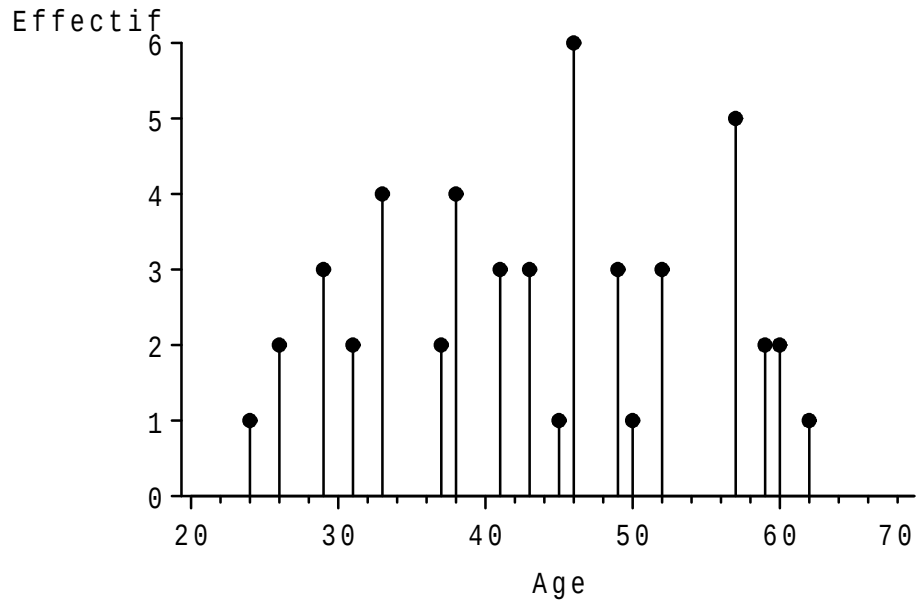


FIG. 2.1 – Diagramme en bâtons

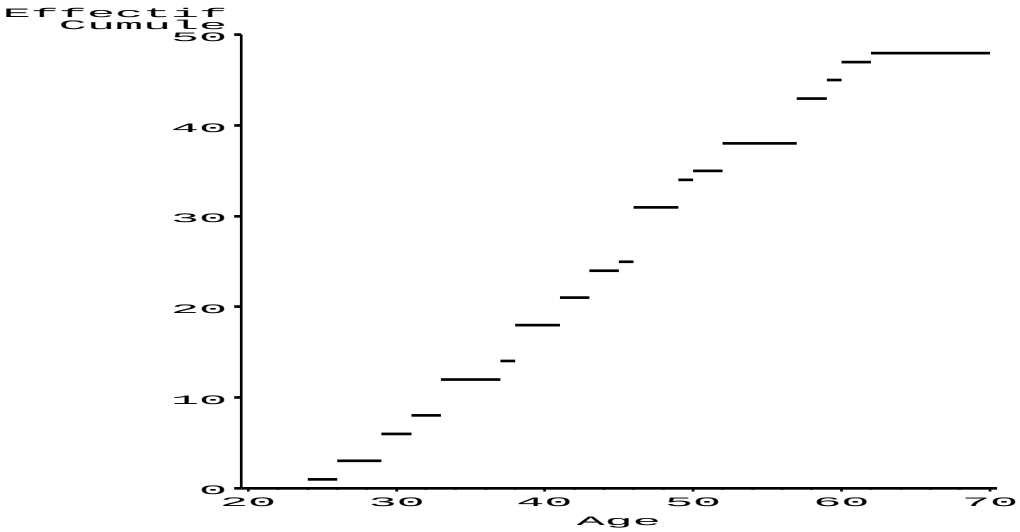


FIG. 2.2 – Diagramme cumulatif

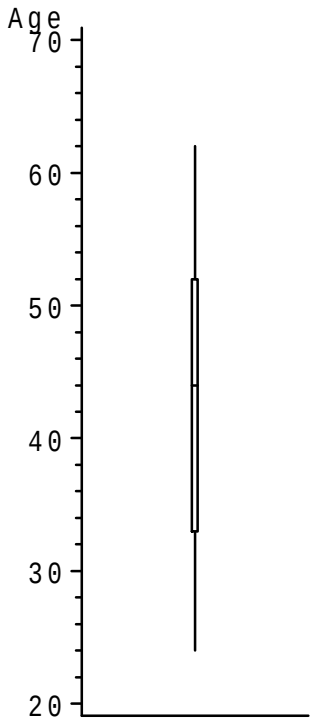


FIG. 2.3 – Boîte–à–moustaches

**Les autres quantiles**

Les *quintiles*, *déciles* et *centiles* sont également d'usage assez courant.

**La boîte–à–moustaches (ou "box–and–whisker plot")**

Il s'agit d'un graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane. La figure 2.3 donne la boîte–à–moustaches de l'exemple 2.1. Dans cet exemple, on a obtenu $x_{\frac{1}{4}} = 35$, $x_{\frac{1}{2}} = 44$ et $x_{\frac{3}{4}} = 52$ ; on notera que l'obtention, d'une part de $x_{\frac{1}{4}}$ et $x_{\frac{1}{2}}$, d'autre part de $x_{\frac{3}{4}}$, ne s'est pas faite de la même façon (en fait, avec une variable discrète, la détermination des quantiles est souvent approximative comme on peut le constater avec cet exemple).

## 1.5 Caractéristiques numériques

Les caractéristiques (ou résumés) numériques introduites ici servent à synthétiser la série étudiée au moyen d'un petit nombre de valeurs numériques. On distingue essentiellement les caractéristiques de tendance centrale (ou encore de *position* ou de *localisation*) et les caractéristiques de dispersion.

**Caractéristiques de tendance centrale**

Leur objectif est de fournir un ordre de grandeur de la série étudiée, c'est–à–dire d'en situer le centre, le milieu. Les deux caractéristiques les plus usuelles sont :

- la *médiane*,
- la *moyenne* (ou moyenne arithmétique).

Formule de la moyenne pour une variable quantitative discrète :

$$\overline{x} \;=\; \frac{1}{n}\sum_{l=1}^{r} n_l x_l \;=\; \sum_{l=1}^{r} f_l x_l .$$

**Caractéristiques de dispersion**

Elles servent à préciser la variabilité de la série, c'est–à–dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

- L'*étendue* $(x_r \;-\; x_1)$,
- l'*intervalle interquartiles* $(x_{\frac{3}{4}} \;-\; x_{\frac{1}{4}})$,
- l'*écart-moyen à la médiane* $(\frac{1}{n}\sum_{l=1}^{r} n_l \left| x_l \;-\; x_{\frac{1}{2}} \right|)$,
- l'*écart-moyen à la moyenne* $(\frac{1}{n}\sum_{l=1}^{r} n_l \left| x_l \;-\; \overline{x} \right|)$,

sont des caractéristiques de dispersion que l'on rencontre parfois.

Mais, la caractéristique de loin la plus utilisée est l'*écart–type*, racine carrée positive de la *variance*. Formules de la variance :

$$\begin{aligned} var(X) = \sigma_X^2 \;&=\; \frac{1}{n}\sum_{l=1}^{r} n_l (x_l \;-\; \overline{x})^2 \\ &=\; \frac{1}{n}\sum_{l=1}^{r} n_l (x_l)^2 \;-\; (\overline{x})^2 . \end{aligned}$$

L'écart–type de $X$ sera donc noté $\sigma_X$.

**Illustration**

En utilisant toujours l'exemple 2.1, on a calculé :

$$\begin{aligned} \overline{x} \;&=\; \frac{1}{n}\sum_{l=1}^{r} n_l x_l \;=\; \frac{2094}{48} \;=\; 43,625 \;\simeq\; 43,6 \text{ ans} ; \\ \sigma_X^2 \;&=\; \frac{1}{n}\sum_{l=1}^{r} n_l (x_l)^2 \;-\; (\overline{x})^2 \;=\; \frac{96620}{48} \;-\; (43,625)^2 \;\simeq\; 109,7760 ; \\ \sigma_X \;&=\; \sqrt{\sigma_X^2} \;\simeq\; 10,5 \text{ ans}. \end{aligned}$$

**Remarques 2.2** *Toutes les caractéristiques numériques introduites ici (médiane, moyenne, variance, écart–type...) sont dites* empiriques, *c'est–à–dire calculées sur un échantillon $\Omega$ ; par opposition, on parle, par exemple, de moyenne* théorique *(ou espérance mathématique) pour désigner le concept de moyenne relatif à une variable aléatoire réelle. De la même manière, toutes les caractéristiques numériques introduites au chapitre 3 (covariance, coefficient de corrélation linéaire...) seront empiriques.*

# 2 Variable quantitative continue

## 2.1 Généralités

Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels. Cela signifie que, dans ce cas, le sous–ensemble de $\mathbb{R}$ des valeurs possibles de la variable étudiée a été divisé en $r$ intervalles contigus appelés *classes*.

En général, les deux raisons principales qui peuvent amener à considérer comme continue une variable quantitative sont le grand nombre d'observations distinctes (un traitement en discret serait dans ce cas peu commode) et le caractère "sensible" d'une variable (il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis). Deux exemples de variables quantitatives fréquemment considérées comme continues sont l'âge et le revenu (pour un groupe d'individus).

Nous noterons $(b_0\,;b_1),\ldots,(b_{r-1}\,;b_r)$ les classes considérées. Les nombres $b_{l-1}$ et $b_l$ sont appelés les *bornes* de la $l^{\text{ième}}$ classe ; $\frac{b_{l-1}+b_l}{2}$ est le *centre* de cette classe et $(b_l - b_{l-1})$ en est l'*amplitude* (en général notée $a_l$).

## 2.2 Présentation des données

On utilise encore un tableau statistique analogue à celui vu au paragraphe précédent, en disposant dans la première colonne les classes rangées par ordre croissant. Les notions d'effectifs, de fréquences, d'effectifs cumulés et de fréquences cumulées sont définies de la même façon que dans le cas discret. On notera que l'on n'utilise pas dans ce cas la présentation tige–et–feuille car les valeurs exactes de la série sont inconnues.

**Exemple 2.2** *Le tableau ci-dessous donne, pour l'année 1987, la répartition des exploitations agricoles françaises selon la SAU (surface agricole utilisée) exprimée en hectares (Tableaux Economiques de Midi–Pyrénées, INSEE, 1989, p. 77) ; la SAU est ici une variable quantitative continue comportant 6 classes.*

| SAU (en ha) | fréquences (%) |
|---|---|
| moins de 5 | 24,0 |
| de 5 à 10 | 10,9 |
| de 10 à 20 | 17,8 |
| de 20 à 35 | 20,3 |
| de 35 à 50 | 10,2 |
| plus de 50 | 16,8 |

## 2.3 Représentations graphiques

Les deux graphiques usuels remplaçant respectivement dans ce cas le diagramme en bâtons et le diagramme cumulatif sont l'histogramme et la courbe cumulative.

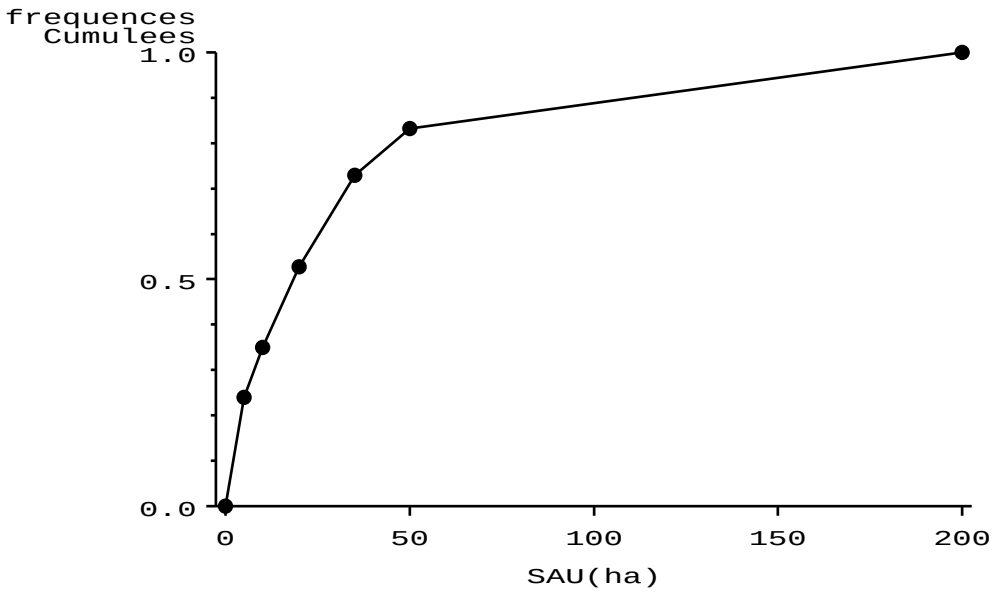


FIG. 2.4 – Courbe cumulative

**La courbe cumulative**

C'est encore une fois le graphe de la *fonction de répartition empirique*, cette dernière devant maintenant être précisée au moyen d'*interpolations linéaires*.

On appelle fonction de répartition empirique de la variable continue $X$ la fonction $F_X$ définie par :

$$F_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ F_{l-1} + \frac{f_l}{b_l - b_{l-1}}(x - b_{l-1}) & \text{si } b_{l-1} \leq x < b_l, \quad l = 1, \ldots, r, \\ 1 & \text{si } x \geq b_r \end{cases}$$

(on a supposé $F_0 = 0$).

La courbe cumulative relative à l'exemple 2.2 est donnée par la figure 2.4. On notera que dans cet exemple, comme c'est souvent le cas avec une variable quantitative continue, il a fallu fixer arbitrairement la borne inférieure de la première classe (il était naturel ici de prendre $b_0 = 0$) ainsi que la borne supérieure de la dernière classe (on a choisi $b_6 = 200$, mais d'autres choix étaient possibles).

**L'histogramme**

La fonction de répartition empirique est, dans le cas continu, une fonction dérivable sauf, éventuellement, aux points d'abscisses $b_0$, $b_1$, ..., $b_r$. Sa fonction dérivée, éventuellement non définie en ces points, est appelée *densité empirique* de $X$ et notée $f_X$. On obtient :

$$f_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ \frac{f_l}{b_l - b_{l-1}} & \text{si } b_{l-1} < x < b_l, \quad l = 1, \ldots, r, \\ 0 & \text{si } x \geq b_r. \end{cases}$$

Le graphe de $f_X$ est alors appelé histogramme de la variable $X$. Un histogramme est donc la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_l = b_l - b_{l-1}$)

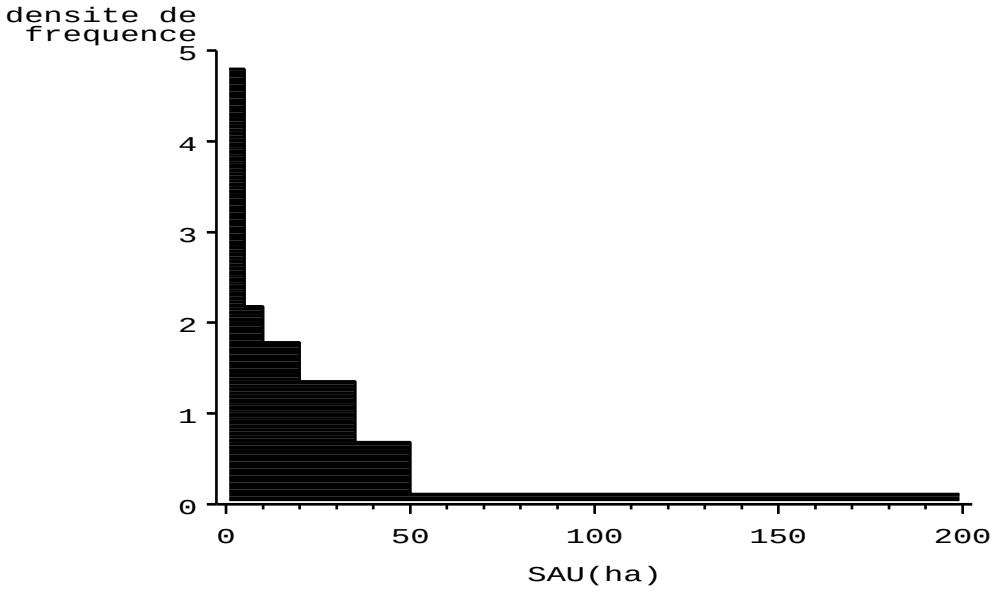


FIG. 2.5 – Histogramme

et dont les hauteurs sont les quantités $\frac{f_l}{b_l - b_{l-1}}$, appelées *densités de fréquence*. L'aire du $l^{\text{ième}}$ rectangle vaut donc $f_l$, fréquence de la classe correspondante.

L'histogramme correspondant aux données de l'exemple 2.2 est présenté dans la figure 2.5.

**Estimation fonctionnelle**

La qualité de l'estimation d'une distribution par un histogramme dépend beaucoup du découpage en classe. Malheureusement, plutôt que de fournir des classes d'effectifs égaux et donc de mieux répartir l'imprécision, les logiciels utilisent des classes d'amplitudes égales et tracent donc des histogrammes parfois peu représentatifs. Ces 20 dernières années, à la suite du développement des moyens de calcul, sont apparues des méthodes d'estimation dites *fonctionnelles* ou *non-paramétriques* qui proposent d'estimer la distribution d'une variable ou la relation entre deux variables par une fonction construite point par point (noyaux) ou dans une base de fonctions *splines*. Ces estimations sont simples à calculer (pour l'ordinateur) mais nécessitent le choix d'un paramètre dit de *lissage*. Les démonstrations du caractère optimal de ces estimations fonctionnelles, liée à l'optimalité du choix de la valeur du paramètre de lissage, font appel à des outils théoriques plus sophistiquées sortant du cadre de ce cours (Eubank, 1988, Silverman, 1986).

L'estimation de la densité par la méthode du noyau se met sous la forme générale :

$$\widehat{g}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right)$$

où $\lambda$ est le paramètre de lissage optimisée par une procédure automatique qui minimise une approximation de l'erreur quadratique moyenne intégrée (norme de l'espace $L^2$) ; $K$ est une fonction symétrique, positive, concave, appelée *noyau* dont la forme précise importe peu. C'est souvent la fonction densité de la loi gaussienne :

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$$

qui possède de bonnes propriétés de régularité. Le principe consiste simplement à associer à chaque observation un "élément de densité" de la forme du noyau $K$ et à sommer tous ces éléments. Un histogramme est une version particulière d'estimation dans laquelle l"'élément de densité" est un "petit rectangle" dans la classe de l'observation.

## 2.4 Détermination des quantiles

Les quantiles $x_\alpha$ d'une variable continue peuvent être déterminés de façon directe à partir de la courbe cumulative. Cela signifie que, par le calcul, on doit commencer par déterminer la classe dans laquelle se trouve le quantile cherché, puis le déterminer dans cette classe par interpolation linéaire (voir l'illustration plus loin).

## 2.5 Détermination des autres caractéristiques numériques

La moyenne, la variance et l'écart–type d'une variable continue se déterminent de la même manière que dans le cas discret ; dans les formules, on doit prendre pour $x_l$ les centres de classes au lieu des observations (qui ne sont pas connues). Les valeurs obtenues pour ces caractéristiques sont donc assez approximatives ; cela n'est pas gênant dans la mesure où le choix de traiter une variable quantitative comme continue correspond à l'acceptation d'une certaine imprécision dans le traitement statistique.

## 2.6 Illustration

La médiane de la variable présentée dans l'exemple 2.2 se situe dans la classe $(10\,;\,20)$, puisque la fréquence cumulée de cette classe (52,7) est la première à dépasser 50. On détermine la médiane en faisant l'interpolation linéaire suivante (l'indice $l$ ci–dessous désigne en fait la troisième classe) :

$$\begin{aligned} x_{\frac{1}{2}} &= b_{l-1} + a_l \frac{50 - F_{l-1}}{F_l - F_{l-1}} \\ &= 10 + 10 \frac{15,1}{17,8} \\ &\simeq 18,5 \text{ ha.} \end{aligned}$$

La moyenne vaut :

$$\overline{x} = \sum_{l=1}^{r} f_l x_l = \frac{3080,5}{100} \simeq 30,8 \text{ ha.}$$

**Remarques 2.3** *Dans cet exemple, il convient de noter trois choses :*

- *tout d'abord, pour le calcul de la moyenne, nous avons choisi $x_6 = 100$, plutôt que 125, car cette valeur nous a semblé plus proche de la réalité ;*
- *ensuite, il se trouve que, dans ce cas, on peut calculer la* vraie *valeur de la moyenne, connaissant la SAU totale en France (31 285 400 ha) et le nombre total d'exploitations agricoles (981 720) ; on obtient 31,9 ha, ce qui signifie que l'approximation obtenue ici est très correcte ;*
- *enfin, le fait que la médiane soit sensiblement plus faible que la moyenne caractérise les séries fortement concentrées sur les petites valeurs.*

# 3 Variable qualitative

## 3.1 Variables nominales et variables ordinales

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées *modalités*. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite *ordinale*. Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite *nominale*.

## 3.2 Traitements statistiques

Il est clair qu'on ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinale). Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques. Encore faut–il noter que les notions d'effectifs cumulés et de fréquences cumulées n'ont de sens que pour des variables ordinales (elles ne sont pas définies pour les variables nominales).

## 3.3 Représentations graphiques

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont :

- le *diagramme en colonnes,*
- le *diagramme en barre,*
- le *diagramme en secteurs.*

Les figures 2.7, 2.6 et 2.8 présentent chacun de ces trois graphiques sur les données de l'exemple 2.3.

**Exemple 2.3** *Le tableau ci–dessous donne la répartition de la population active occupée (ayant effectivement un emploi) selon la CSP (catégorie socioprofessionnelle), en France, en mars 1988 (Tableaux de l'Economie Française, INSEE, 1989, p. 59).*

| CSP | effectifs en milliers | fréquences (%) |
|---|---|---|
| 1. agriculteurs exploitants | 1312 | 6,1 |
| 2. artisans, commerçants, chefs d'entreprises | 1739 | 8,1 |
| 3. cadres, professions intellectuelles supérieures | 2267 | 10,6 |
| 4. professions intermédiaires | 4327 | 20,1 |
| 5. employés | 5815 | 27,0 |
| 6. ouvriers | 6049 | 28,1 |

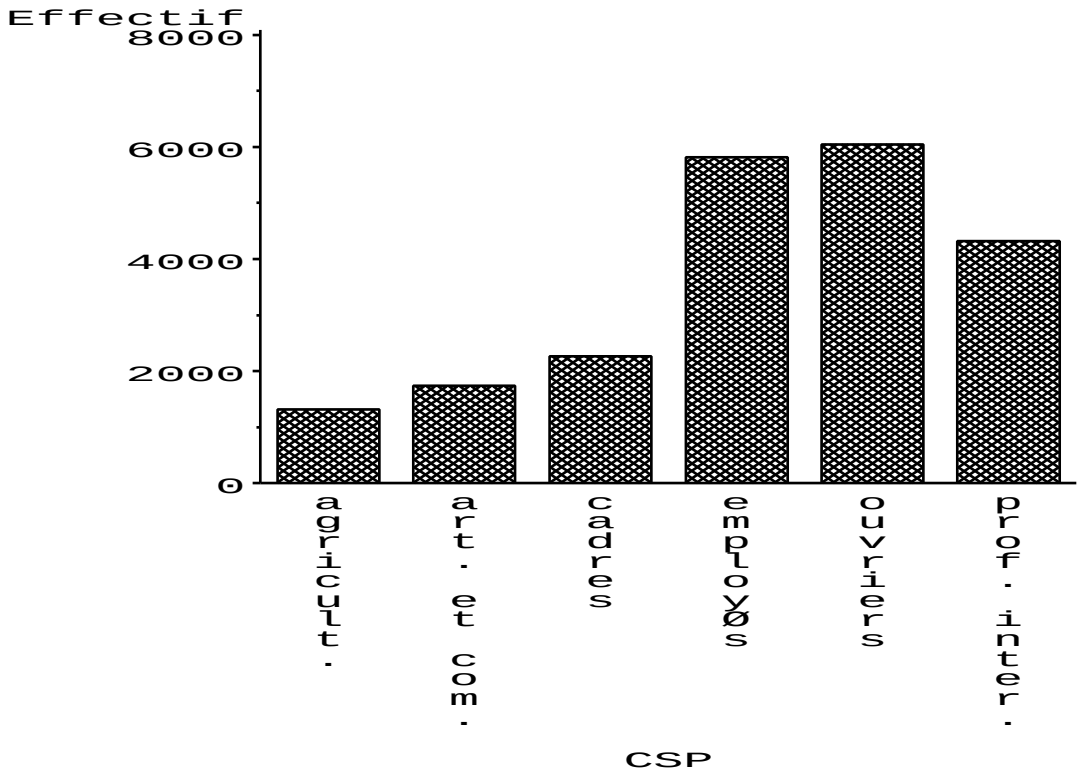


FIG. 2.6 – Diagramme en colonnes


FIG. 2.7 – Diagramme en barre

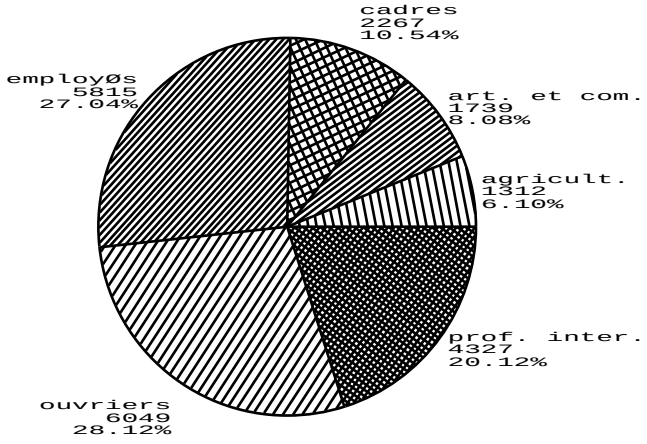


FIG. 2.8 – Diagramme en secteurs

# Chapitre 3

# Cas bidimensionnel

Dans tout ce chapitre, on s'intéresse à l'étude simultanée de deux variables, notées $X$ et $Y$, étudiées sur le même échantillon, toujours noté $\Omega$. L'objectif essentiel des méthodes présentées est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors *liaison*. Dans certains cas, cette liaison peut être considérée a priori comme *causale*, une variable expliquant l'autre ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations. Sont ainsi introduites les notions de covariance, coefficient de corrélation linéaire, régression linéaire, rapport de corrélation, khi–deux et autres indicateurs qui lui sont liés. De même, nous présentons les graphiques illustrant les liaisons entre variables : nuage de points (*scatter–plot*), boîtes–à–moustaches parallèles, diagramme de profils, tableau de nuages (*scatter–plot matrix*).

## 1 Deux variables quantitatives

### 1.1 Les données

Les variables quantitatives considérées dans ce chapitre seront toujours discrètes ; en effet, cela n'a pas d'intérêt de regrouper les valeurs identiques de l'une des deux variables, puisqu'elles ne correspondent pas nécessairement à des valeurs identiques de l'autre. La donnée de base est donc constituée par la série statistique brute se présentant sous la forme $\{(X(\omega_i), Y(\omega_i)) \, ; i = 1, \ldots, n\}$.

**Exemple 3.1** *Les données ci–dessous seront utilisées tout au long de ce chapitre ; elles sont extraites des Tableaux de l'Economie Française, INSEE, 1989, p.109. Pour 51 secteurs d'activité (numérotés de 04 à 54, selon la nomenclature NAP 100), on a considéré (au 01.01.1986, en France) le nombre total d'entreprises (variable notée $NB$), l'effectif salarié (notée $EF$), et le chiffre d'affaires hors–taxes en millions de francs (notée $CA$). Dans le tableau ci–dessous, nous donnons les trois premières et les trois dernières lignes de l'ensemble des données.*

| code | secteur | $NB$ | $EF$ | $CA$ |
|---|---|---|---|---|
| 04 | production de combustibles minéraux solides et cokéfaction | 19 | 49251 | 14111 |
| 05 | production de pétrole et de gaz naturel | 120 | 46594 | 306293 |
| 06 | production et distribution d'electricité | 731 | 129723 | 138389 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 52 | industrie du caoutchouc | 746 | 87121 | 37502 |
| 53 | transformation des matières plastiques | 3232 | 102437 | 58122 |
| 54 | industries diverses | 10171 | 84012 | 38071 |

Pour mémoire, nous indiquons ci–dessous quelques caractéristiques numériques des trois variables considérées.

| variable | minimum | maximum | moyenne | écart–type |
|---|---|---|---|---|
| $NB$ | 11 | 41866 | 4135 | 7435 |
| $EF$ | 1701 | 425082 | 92591 | 83832 |
| $CA$ | 992 | 306293 | 68010 | 64532 |

## 1.2 Représentation graphique : le nuage de points

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable $X$ et l'axe vertical la variable $Y$, puis à représenter chaque individu observé $\omega_i$ par le point d'abscisse $X(\omega_i)$ et d'ordonnée $Y(\omega_i)$. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables et est appelé *nuage*. On notera qu'on rencontre parfois la terminologie de *diagramme de dispersion*, traduction fidèle du terme anglais *scatter–plot*.

La figure 3.1 présente le nuage de points réalisé, avec les données de l'exemple 3.1, en utilisant les variables $CA$ (en ordonnées) et $EF$ (en abscisses). De plus, on a tracé la droite de régression de $CA$ sur $EF$ (voir le paragraphe 1.4).

**Remarques 3.1** *Le choix des échelles à retenir pour réaliser un nuage de points peut s'avérer délicat. D'une façon générale, on distinguera le cas de variables* homogènes *(représentant la même grandeur et exprimées dans la même unité) de celui des variables* hétérogènes. *Dans le premier cas, on choisira la même échelle sur les deux axes (qui seront donc orthonormés) ; dans le second cas, il est recommandé soit de représenter les variables centrées et réduites (voir ci–dessous) sur des axes orthonormés, soit de choisir des échelles telles que ce soit sensiblement ces variables là que l'on représente (c'est en général cette seconde solution qu'utilisent, de façon automatique, les logiciels statistiques).*

### Rappel : variables centrées et réduites

Si $X$ est une variable quantitative de moyenne $\overline{x}$ et d'écart–type $\sigma_X$, on appelle variable centrée associée à $X$ la variable $X - \overline{x}$ (elle est de moyenne nulle et d'écart–type $\sigma_X$), et variable centrée et réduite (ou tout simplement variable réduite) associée à $X$ la variable $\frac{X - \overline{x}}{\sigma_X}$ (elle est de moyenne nulle et d'écart–type égal à un). Une variable centrée et réduite s'exprime sans unité.

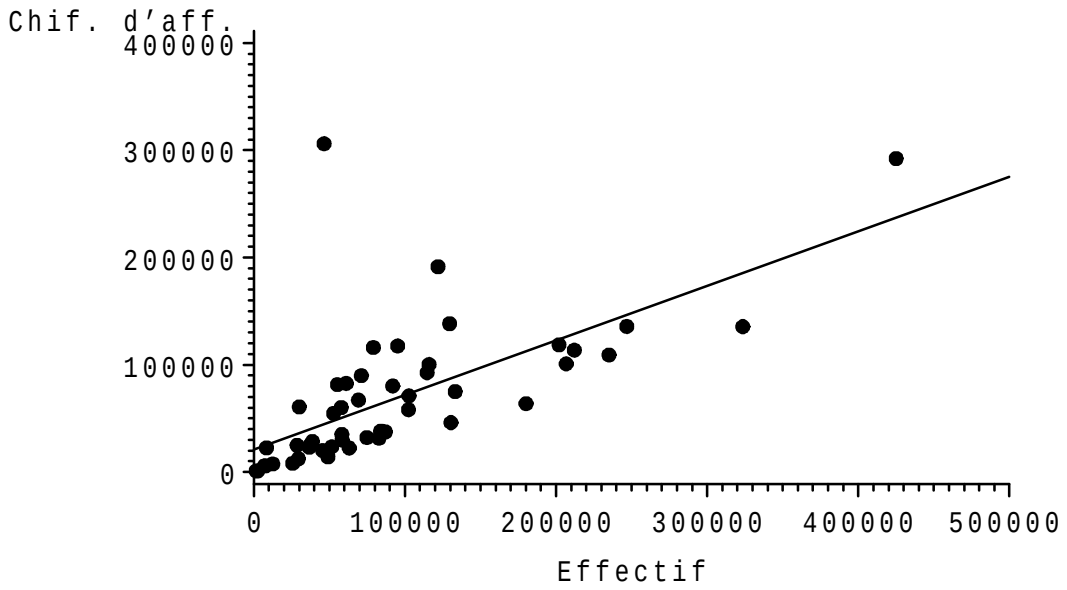


FIG. 3.1 – Nuage de points

## 1.3 La covariance et le coefficient de corrélation linéaire

L'objectif est maintenant de définir un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Cet indice est le coefficient de corrélation linéaire ; il nécessite la définition préalable de la covariance.

**La covariance : définition**

La covariance généralise à deux variables la notion de variance ; sa formule de définition est la suivante :

$$\begin{aligned} cov(X,Y) = c_{XY} &= \frac{1}{n}\sum_{i=1}^{n}[X(\omega_i)-\overline{x}][Y(\omega_i)-\overline{y}] \\ &= [\frac{1}{n}\sum_{i=1}^{n}X(\omega_i)Y(\omega_i)]-\overline{x}\ \overline{y}. \end{aligned}$$

**La covariance : propriétés**

- La covariance est un indice *symétrique*. De façon évidente, on a $c_{XY} = c_{YX}$ (les deux variables jouent donc le même rôle dans la définition de la covariance).
- La covariance peut prendre *toute valeur réelle* (négative, nulle ou positive).
- La covariance est une *forme bilinéaire symétrique* dont la variance est la *forme quadratique* associée. En particulier, on en déduit les deux formules suivantes :

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y),$$

$$[cov(X,Y)]^2 \leq var(X)var(Y)\,;$$

(cette dernière propriété est l'inégalité de Cauchy-Schwarz).

**Le coefficient de corrélation linéaire: définition**

Il est clair que la covariance dépend des unités de mesure dans lesquelles sont exprimées les variables considérées ; en ce sens, ce n'est pas un indice de liaison "intrinsèque". C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (parfois appelé coefficient de Pearson ou de Bravais–Pearson), rapport entre la covariance et le produit des écarts–types :

$$corr(X,Y) = r_{XY} = \frac{c_{XY}}{\sigma_X \sigma_Y}.$$

**Le coefficient de corrélation linéaire: propriétés**

- Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à $X$ et $Y$ : $r_{XY} = cov(\frac{X-\overline{x}}{\sigma_X}, \frac{Y-\overline{y}}{\sigma_Y})$. Par conséquent, $r_{XY}$ est indépendant des unités de mesure de $X$ et de $Y$.
- Le coefficient de corrélation est *symétrique*: $r_{XY} = r_{YX}$ .
- $-1 \leq r_{XY} \leq +1$ . Les valeurs $-1$ et $+1$ correspondent à une liaison linéaire parfaite entre $X$ et $Y$ (existence de réels $a$, $b$ et $c$ tels que : $aX + bY + c = 0$).

**Illustration**

En reprenant les données de l'exemple 3.1, nous avons calculé la covariance et le coefficient de corrélation linéaire entre les variables $EF$ et $CA$ ; on a obtenu :

$$cov(EF,CA) = 35769 \times 10^5 \ ; \ corr(EF,CA) = \frac{35769 \times 10^5}{83832 \times 64532} \simeq 0,66 \ .$$

La liaison linéaire entre les deux variables est donc positive et moyenne.

## 1.4 Regression linéaire entre deux variables

**Introduction**

Lorsque deux variables quantitatives sont correctement correlées ($|r_{XY}|$ voisin de 1), et que l'on peut a priori considérer que l'une (nous supposerons qu'il s'agit de $X$) est cause de l'autre (il s'agira donc de $Y$), il est naturel de chercher, dans un ensemble donné de fonctions, la fonction de $X$ approchant $Y$ "le mieux possible", au sens d'un certain critère ; on dit que l'on fait la regression de $Y$ sur $X$. Si l'on choisit pour ensemble de fonctions celui des fonctions affines (du type $aX+b$), on parle alors de régression linéaire. C'est le choix que l'on fait le plus fréquemment dans la pratique, le critère le plus usuel étant celui des moindres carrés.

**Le critère des moindres carrés**

Il consiste à minimiser la quantité suivante :

$$S(a,b) = \sum_{i=1}^{n} \{Y(\omega_i) - [aX(\omega_i) + b]\}^2.$$

On notera que $|Y(\omega_i) - [aX(\omega_i) + b]|$ représente, dans le nuage de points, la distance verticale du point figurant $\omega_i$ à la droite d'équation $y = ax + b$.

**Solution**

La minimisation de $S$ en $a$ et $b$ fournit la solution unique suivante :

$$\widehat{a} = \frac{c_{XY}}{\sigma_X^2} \ ; \ \widehat{b} = \overline{y} - \widehat{a}\overline{x}.$$

**Propriétés**

- La droite d'équation $y = \widehat{a}x + \widehat{b}$ est appelée *droite de régression* de $Y$ sur $X$ ; elle passe par le *barycentre* du nuage des $\omega_i$, de coordonnées $(\overline{x}, \overline{y})$.
- Les valeurs $\widehat{y_i} = \widehat{a}X(\omega_i) + \widehat{b}$ sont appelées les *valeurs ajustées* ; elles ont la même moyenne $\overline{y}$ que $Y$.
- Les valeurs $\widehat{e_i} = Y(\omega_i) - \widehat{y_i}$ sont appelées les *résidus*. Ils sont de moyenne nulle et de variance $\frac{1}{n}S(\widehat{a}, \widehat{b})$.
- La variable causale $X$ et la variable résiduelle $\widehat{E}$ sont non correlées :

$$corr(X, \widehat{E}) = 0.$$

**Illustration**

En considérant toujours l'exemple 3.1, nous avons réalisé la régression linéaire de la variable $CA$ sur la variable $EF$. On a obtenu :

$$\hat{a} = \frac{cov(EF, CA)}{var(EF)} = \frac{35769 \times 10^5}{70278 \times 10^5} \simeq 0,509 \ ;$$

$$\hat{b} = \overline{CA} \ - \ \hat{a} \ \overline{EF} = 68010 - 0,509 \times 92591 \simeq 20884 \ .$$

La droite de régression correspondante a été tracée sur la figure 3.1.

# 2 Une variable quantitative et une qualitative

## 2.1 Les données

Soit $X$ la variable qualitative considérée, supposée à $r$ modalités notées

$$x_1, \ldots, x_l, \ldots, x_r$$

et soit $Y$ la variable quantitative de moyenne $\overline{y}$ et de variance $\sigma_Y^2$. Désignant toujours par $\Omega$ l'échantillon considéré, chaque modalité $x_l$ de $X$ définit une sous–population (un sous–ensemble) $\Omega_l$ de $\Omega$ : c'est l'ensemble des individus sur lesquels on a observé $x_l$ ; on obtient ainsi une *partition* de $\Omega$ en $r$ classes dont nous noterons $n_1, \ldots, n_r$ les cardinaux (avec toujours $\sum_{l=1}^{r} n_l = n$, où $n = card(\Omega)$).

Considérant alors la restriction de $Y$ à $\Omega_l$ $(l = 1, \ldots, r)$, on peut définir la moyenne et la variance partielles de $Y$ sur cette sous–population ; nous les noterons respectivement $\overline{y_l}$ et $\sigma_l^2$ :

$$\overline{y_l} = \frac{1}{n_l} \sum_{\omega_i \in \Omega_l} Y(\omega_i) \, ;$$

$$\sigma_l^2 = \frac{1}{n_l} \sum_{\omega_i \in \Omega_l} [Y(\omega_i) - \overline{y_l}]^2.$$

**Exemple 3.2** *Cet exemple est présenté dans Johnson & Wichern (1988), page 218 ; les individus sont 19 chiens prémédiqués au pentobarbital, et l'on étudie l'effet sur leur rythme cardiaque de deux facteurs (variables qualitatives explicatives) croisés. L'effet est mesuré par le temps entre deux battements de cœur successifs (variable quantitative $Y$, mesurée en millisecondes), et les deux facteurs, à deux niveaux chacun, sont la pression d'administration de dioxyde de carbone ($CO_2$), qui peut être élevée (E) ou faible (F), et la présence (1) ou l'abscence (0) d'halothane ; la variable $X$ est celle obtenue par le croisement de ces deux facteurs ; elle est donc qualitative à 4 modalités : $x_1 = E0, x_2 = F0, x_3 = E1, x_4 = F1$. L'expérience ayant été répétée 4 fois sur chaque chien (une fois dans chacune des 4 conditions ainsi définies), on dispose donc de $n = 4 \times 19 = 76$ individus. Les données se trouvent dans le tableau ci-dessous.*

| | modalité de X | | | |
|---|---|---|---|---|
| numéro du chien | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
| 1 | 426 | 609 | 556 | 600 |
| 2 | 253 | 236 | 392 | 395 |
| 3 | 359 | 433 | 349 | 357 |
| 4 | 432 | 431 | 522 | 600 |
| 5 | 405 | 426 | 513 | 513 |
| 6 | 324 | 438 | 507 | 539 |
| 7 | 310 | 312 | 410 | 456 |
| 8 | 326 | 326 | 350 | 504 |
| 9 | 375 | 447 | 547 | 548 |
| 10 | 286 | 286 | 403 | 422 |
| 11 | 349 | 382 | 473 | 497 |
| 12 | 429 | 410 | 488 | 547 |
| 13 | 348 | 377 | 447 | 514 |
| 14 | 412 | 473 | 472 | 446 |
| 15 | 347 | 326 | 455 | 468 |
| 16 | 434 | 458 | 637 | 524 |
| 17 | 364 | 367 | 432 | 469 |
| 18 | 420 | 395 | 508 | 531 |
| 19 | 397 | 556 | 645 | 625 |

Nous indiquons également les moyennes et les écarts–types partiels des 4 sous–populations, ainsi que la moyenne et l'écart–type de la population globale.

| | $\Omega_1$ | $\Omega_2$ | $\Omega_3$ | $\Omega_4$ | $\Omega$ |
|---|---|---|---|---|---|
| moyennes | 368, 2 | 404, 6 | 479, 3 | 502, 9 | 438, 8 |
| écarts–types | 51, 7 | 86, 9 | 80, 6 | 68, 0 | 91, 1 |

**Remarques 3.2** *Ces données sont un peu particulières, dans la mesure où d'une part les individus sont dupliqués 4 fois, et d'autre part les 4 modalités de la variable $X$ correspondent au croisement de deux facteurs. Ainsi, d'autres traitements statistiques que ceux indiqués ici sont envisageables (voir Johnson & Wichern, 1988).*

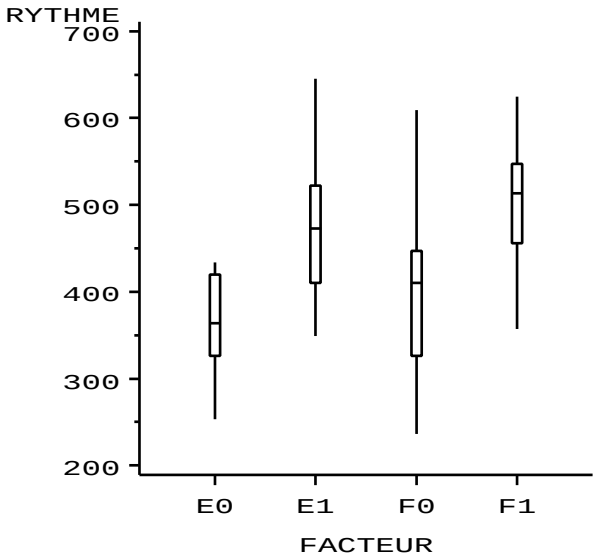


FIG. 3.2 – Boîtes parallèles

## 2.2 Représentation graphique : les boîtes parallèles

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des boîtes parallèles ; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour $Y$ une boîte–à–moustaches pour chacune des sous–populations définies par $X$. La comparaison de ces boîtes donne une idée assez claire de l'influence de $X$ sur les valeurs de $Y$, c'est–à–dire de la liaison entre les deux variables. La figure 3.2 donne les boîtes parallèles de l'exemple 3.2.

## 2.3 Formules de décomposition

Ces formules indiquent comment se décomposent la moyenne et la variance de $Y$ sur la partition définie par $X$ (c'est–a–dire comment s'écrivent ces caractéristiques en fonction de leurs valeurs partielles) ; elles sont nécessaires pour définir un indice de liaison entre les deux variables. Ces formules sont les suivantes :

$$\begin{aligned} \overline{y} &= \frac{1}{n}\sum_{l=1}^{r} n_l \overline{y_l}\,; \\ \sigma_Y^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l (\overline{y_l} - \overline{y})^2 \;+\; \frac{1}{n}\sum_{l=1}^{r} n_l \sigma_l^2 = \sigma_E^2 + \sigma_R^2 \;. \end{aligned}$$

Le premier terme de la décomposition de $\sigma_Y^2$, noté $\sigma_E^2$, est appelé *variance expliquée* (par la partition, c'est–à–dire par $X$) ; le second terme, noté $\sigma_R^2$, est appelé *variance résiduelle*.

On notera qu'une formule de décomposition analogue existe pour la covariance entre deux variables quantitatives.

## 2.4 Le rapport de correlation

**Définition**

Il s'agit d'un indice de liaison entre les deux variables $X$ et $Y$ ; il est défini de la façon suivante :

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}} \ .$$

**Propriétés**

- $s_{Y/X}$ n'est pas symétrique. Cette propriété est évidente, compte–tenu que $X$ et $Y$ ne sont pas de même nature.
- $0 \leq s_{Y/X} \leq 1$. Cet encadrement de $s_{Y/X}$ découle directement de la formule de décomposition de la variance. Les valeurs 0 et 1 ont une signification particulière intéressante.

**Illustration**

Sur l'exemple 3.2 la variance totale vaut 8305,90 , la variance expliquée 2973,94 et la variance résiduelle 5331,96. On en déduit que le rapport de corrélation vaut :

$$s_{Y/X} = \sqrt{\frac{2973,94}{8305,90}} \simeq 0,60.$$

La liaison entre $X$ et $Y$ est donc moyenne.

# 3 Deux variables qualitatives

## 3.1 Les données et leur présentation

On considère dans ce paragraphe deux variables qualitatives observées simultanément sur $n$ individus. On suppose que la première, notée $X$, possède $r$ modalités notées $x_1, \ldots, x_l, \ldots, x_r$, et que la seconde, notée $Y$, possède $c$ modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

Le plus souvent, ces données sont présentées dans un tableau à double entrée, appelé *table de contingence*, dans lequel on dispose les modalités de $X$ en lignes et celles de $Y$ en colonnes. Ce tableau est donc de dimension $r \times c$ et a pour élément générique le nombre $n_{lh}$ d'observations conjointes des modalités $x_l$ de $X$ et $y_h$ de $Y$ ; les quantités $n_{lh}$ sont appelées les *effectifs conjoints*.

Une table de contingence se présente donc sous la forme suivante :

| | $y_1$ | $\cdots$ | $y_h$ | $\cdots$ | $y_c$ | sommes |
|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $\cdots$ | $n_{1h}$ | $\cdots$ | $n_{1c}$ | $n_{1+}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $x_l$ | $n_{l1}$ | $\cdots$ | $n_{lh}$ | $\cdots$ | $n_{lc}$ | $n_{l+}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $x_r$ | $n_{r1}$ | $\cdots$ | $n_{rh}$ | $\cdots$ | $n_{rc}$ | $n_{r+}$ |
| sommes | $n_{+1}$ | $\cdots$ | $n_{+h}$ | $\cdots$ | $n_{+c}$ | $n$ |

Les quantités $n_{l+}(l = 1, \ldots, r)$ et $n_{+h}(h = 1, \ldots, c)$ sont appelées les *effectifs marginaux*; ils sont définis par $n_{l+} = \sum_{h=1}^{c} n_{lh}$ et $n_{+h} = \sum_{l=1}^{r} n_{lh}$, et ils vérifient $\sum_{l=1}^{r} n_{l+} = \sum_{h=1}^{c} n_{+h} = n$.

De façon analogue, on peut définir les notions de fréquences conjointes et de fréquences marginales.

**Exemple 3.3** *Dans cet exemple, on a considéré un échantillon de 797 étudiants de l'Université Paul Sabatier (Toulouse) ayant obtenu soit le DEUG A soit le DEUG B (diplômes scientifiques de premier cycle), et uniquement ce diplôme, durant la période 1971–1983. Quatre variables ont été prises en compte, toutes qualitatives: la série de bac, à 2 modalités (C ou E, D), la mention au bac, à 4 modalités (très bien ou bien, assez bien, passable, inconnue), l'âge d'obtention du bac, à 4 modalités (moins de 18 ans, 18 ans, 19 ans, plus de 19 ans), et la durée d'obtention du DEUG, à 3 modalités (2 ans, 3 ans, 4 ans). Dans la table de contingence ci–dessous, on a croisé en lignes la durée d'obtention du DEUG (variable $X$, à $r = 3$ modalités), et en colonnes l'âge d'obtention du bac (variable $Y$, à $c = 4$ modalités).*

| | $<$ 18 ans | 18 ans | 19 ans | $>$ 19 ans | sommes |
|---|---|---|---|---|---|
| 2 ans | 84 | 224 | 73 | 19 | 400 |
| 3 ans | 35 | 137 | 75 | 27 | 274 |
| 4 ans | 14 | 59 | 34 | 16 | 123 |
| sommes | 133 | 420 | 182 | 62 | 797 |

## 3.2 Les représentations graphiques

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'*adapter* les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas et que l'on appelle les *profils*.

### Définition des profils

On appelle $l^{\text{ième}}$ profil–ligne l'ensemble des fréquences de la variable $Y$ conditionnelles à la modalité $x_l$ de $X$ (c'est–à–dire définies au sein de la sous–population $\Omega_l$ de $\Omega$ associée à cette modalité). Il s'agit donc des quantités :

$$\{\frac{n_{l1}}{n_{l+}}, \ldots, \frac{n_{lh}}{n_{l+}}, \ldots, \frac{n_{lc}}{n_{l+}}\}.$$

On définit de façon analogue le $h^{\text{ième}}$ profil–colonne :

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{lh}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\}.$$

La représentation graphique des profils–lignes ou des profils–colonnes, au moyen, par exemple, de diagrammes en barre parallèles, donne alors une idée assez précise de la variation conjointe des deux variables.

### Illustration

Nous avons déterminé les profils–colonnes (en pourcentages) relatifs à la table de contingence présentée dans l'exemple 3.3 et nous les donnons ci-dessous.

| | $<$ 18 ans | 18 ans | 19 ans | $>$ 19 ans | profil moyen |
|---|---|---|---|---|---|
| 2 ans | 63,2 | 53,3 | 40,1 | 30,6 | 50,2 |
| 3 ans | 26,3 | 32,6 | 41,2 | 43,6 | 34,4 |
| 4 ans | 10,5 | 14,1 | 18,7 | 25,8 | 15,4 |
| sommes | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |

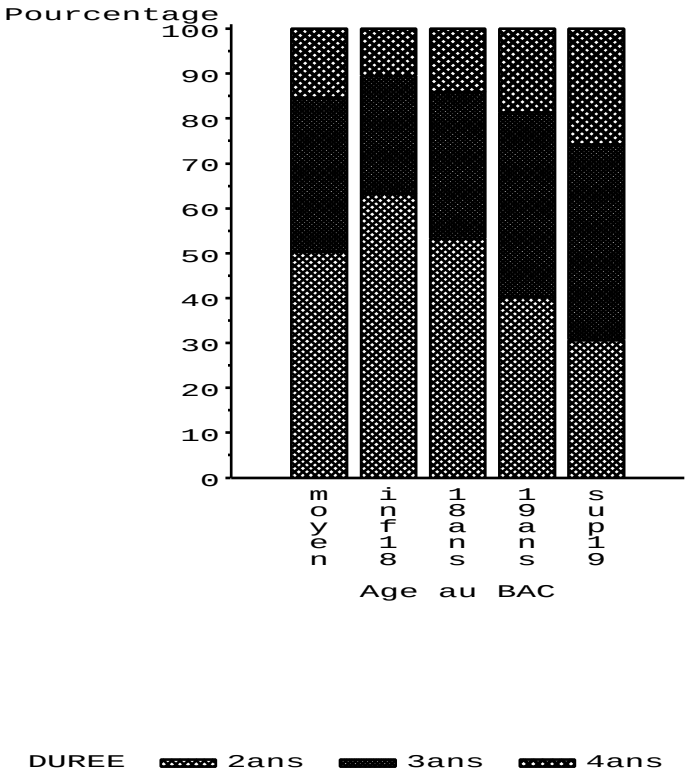


FIG. 3.3 – Diagramme en barres

La figure 3.3 donne le diagramme en barres pour les profils–colonnes ci–dessus, et une liaison entre les deux variables étudiées apparaît très clairement.

## 3.3 Les indices de liaison : le khi–deux et ses dérivés

**Propriété préliminaire**

On peut établir l'équivalence des trois propriétés suivantes :

i. Tous les profils–lignes sont égaux.

ii. Tous les profils–colonnes sont égaux.

iii. $\forall (l,h) \in \{1,\ldots,r\} \times \{1,\ldots,c\}$ :

$$n_{lh} = \frac{n_{l+} n_{+h}}{n} \ .$$

Si une table de contingence vérifie ces trois propriétés, on peut alors dire qu'il n'existe aucune forme de liaison entre les deux variables considérées $X$ et $Y$. Par suite, la mesure de la liaison va se faire en évaluant l'écart entre la situation observée et l'état de non liaison que nous définirons par (iii).

**Définition du khi–deux**

Il est courant en statistique de comparer une table de contingence observée, d'effectif conjoint générique $n_{lh}$, à une table de contingence donnée a priori (et appelée *standard*), d'effectif conjoint générique $s_{lh}$, en calculant la quantité

$$\sum_{l=1}^{r}\sum_{h=1}^{c}\frac{(n_{lh}-s_{lh})^2}{s_{lh}}\ .$$

De façon naturelle, pour mesurer la liaison sur une table de contingence, on utilise donc l'indice appelé khi–deux (chi–square) et défini comme suit :

$$\chi^2=\sum_{l=1}^{r}\sum_{h=1}^{c}\frac{(n_{lh}-\frac{n_{l+}n_{+h}}{n})^2}{\frac{n_{l+}n_{+h}}{n}}=n\left[\sum_{l=1}^{r}\sum_{h=1}^{c}\frac{n_{lh}^2}{n_{l+}n_{+h}}-1\right].$$

Le coefficient $\chi^2$ est toujours positif ou nul et il est d'autant plus grand que la liaison entre les deux variables considérées est forte. Malheureusement, il dépend aussi des dimensions $r$ et $c$ de la table étudiée, ainsi que de la taille $n$ de l'échantillon observé ; en particulier, il n'est pas majoré. C'est la raison pour laquelle on a défini d'autres indices, liés au khi–deux, et dont l'objectif est de palier ces défauts.

**Autres indicateurs liés au khi–deux**

Nous en citerons trois.

- Le *phi–deux* : $\Phi^2=\frac{\chi^2}{n}$. Il ne dépend plus de $n$, mais dépend encore de $r$ et de $c$.
- Le coefficient $T$ de Tschuprow :

  $$T=\sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}.$$

  On peut vérifier : $0\leq T\leq 1$ .
- Le coefficient C de Cramer :

  $$C=\sqrt{\frac{\Phi^2}{d-1}},$$

  avec : $d=\inf(r,c)$. On vérifie maintenant : $0\leq T\leq C\leq 1$ .

**Illustration**

En utilisant les données de l'exemple 3.3, on a obtenu :

$$\chi^2=28,7\,;\ \Phi^2=0,036\,;\ T=0,12\,;\ C=0,13.$$

On a en fait affaire à une liaison peu importante.

# 4 Vers le cas multidimensionnel

L'objectif des prochains chapitres de ce cours est d'exposer les techniques de la statistique descriptive multidimensionnelle. Or, sans connaître ces techniques, il se trouve qu'il est possible de débuter une exploration de données multidimensionnelles en adaptant simplement les méthodes déjà étudiées. C'est cette idée que nous développons dans ce paragraphe.

## 4.1 Les matrices des variances–covariances et des correlations

Lorsqu'on a observé simultanément plusieurs variables quantitatives ($p$ variables, $p \geq 3$) sur le même échantillon, il est possible de calculer d'une part les variances de toutes ces variables, d'autre part les $\frac{p(p-1)}{2}$ covariances des variables prises deux à deux. L'ensemble de ces quantités peut alors être disposé dans une matrice carrée ($p \times p$) et symétrique, comportant les variances sur la diagonale et les covariances à l'extérieur de la diagonale ; cette matrice, appelée matrice des variances–covariances (ou encore matrice des covariances) sera notée **S**. Elle sera utilisée par la suite, mais n'a pas d'interprétation concrète. Notons qu'il est possible de vérifier que **S** est semi définie positive.

De la même manière, on peut construire la matrice symétrique $p \times p$, comportant des 1 sur toute la diagonale et, en dehors de la diagonale, les coefficients de corrélation linéaire entre les variables prises deux à deux. Cette matrice est appelée matrice des corrélations, elle est également semi définie positive, et nous la noterons **R**.. Elle est de lecture commode et indique quelle est la structure de corrélation des variables étudiées.

### Illustration

Nous avons repris ici encore l'exemple 3.1 et calculé les matrices des variances–covariances et des corrélations entre les trois variables intervenant.

$$\mathbf{S} = \begin{bmatrix} 55284 \times 10^3 & 25084 \times 10^4 & 25156 \times 10^3 \\ 25084 \times 10^4 & 70278 \times 10^5 & 35769 \times 10^5 \\ 25156 \times 10^3 & 35769 \times 10^5 & 41644 \times 10^5 \end{bmatrix};$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0,402 & 0,052 \\ 0,402 & 1 & 0,661 \\ 0,052 & 0,661 & 1 \end{bmatrix}.$$

## 4.2 Les tableaux de nuages (ou graphiques splom)

Notons $X^1, \ldots, X^p$ les $p$ variables quantitatives considérées ; on appelle tableau de nuages le graphique obtenu en juxtaposant, dans une sorte de matrice carrée $p \times p$, $p^2$ sous–graphiques ; chacun des sous-graphiques diagonaux est relatif à l'une des $p$ variables, et il peut s'agir, par exemple, d'un histogramme; le sous–graphique figurant dans le bloc d'indice $(j, j'), j \neq j'$, est le nuage de points réalisé avec la variable $X^j$ en abscisses et la variable $X^{j'}$ en ordonnées. Dans certains logiciels anglo–saxons, ces graphiques sont appelés *splom* (Scatter PLOt Matrix). On notera qu'il est possible de ne représenter que la partie supérieure (ou inférieure) du tableau en question. Il est également possible de faire apparaître dans chaque sous–graphique non diagonal une droite de régression. Le tableau de nuages, avec la matrice des corrélations, fournit ainsi une vision globale des liaisons entre les variables étudiées.

### Illustration

La figure 3.4 présente le tableau de nuages relatif aux données de l'exemple 3.1.

## 4.3 La matrice des coefficients de Tschuprow (ou de Cramer)

Considérons maintenant le cas où l'on étudie simultanément plusieurs variables qualitatives ($p$ variables, $p \geq 3$). La matrice des coefficients de Tschuprow est la matrice carrée d'ordre $p$, symétrique, comportant des 1 sur la diagonale et, en dehors de la diagonale, les coefficients de

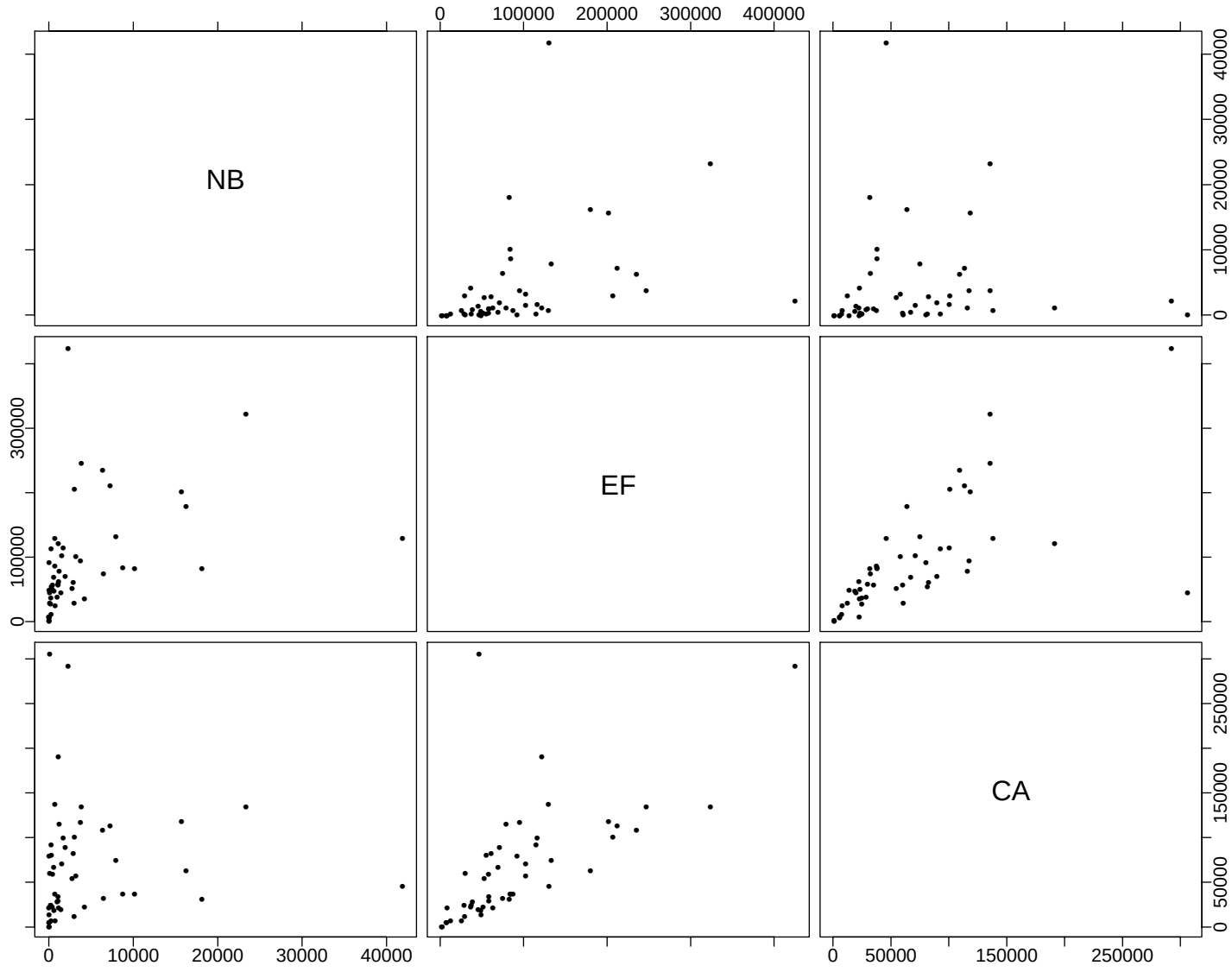


FIG. 3.4 – Tableau des nuages

| | bacC | bacD | < 18 | 18ans | 19ans | > 19 | 2ans | 3ans | 4ans |
|---|---|---|---|---|---|---|---|---|---|
| bacC | 583 | 0 | 108 | 323 | 114 | 38 | 324 | 192 | 67 |
| bacD | 0 | 214 | 25 | 97 | 68 | 24 | 76 | 82 | 56 |
| < 18 | 108 | 25 | 133 | 0 | 0 | 0 | 84 | 35 | 14 |
| 18ans | 323 | 97 | 0 | 420 | 0 | 0 | 224 | 137 | 59 |
| 19ans | 114 | 68 | 0 | 0 | 182 | 0 | 73 | 75 | 34 |
| > 19 | 38 | 24 | 0 | 0 | 0 | 62 | 19 | 27 | 16 |
| 2ans | 324 | 76 | 84 | 224 | 73 | 19 | 400 | 0 | 0 |
| 3ans | 192 | 82 | 35 | 137 | 75 | 27 | 0 | 274 | 0 |
| 4ans | 67 | 56 | 14 | 59 | 34 | 16 | 0 | 0 | 123 |

TAB. 3.1 – Tableau de Burt

Tschuprow entre les variables prises deux à deux. Il s'agit donc d'une matrice du même type que la matrice des corrélations (elle est d'ailleurs, elle aussi, semi définie positive), et son utilisation pratique est analogue. Notons que l'on peut, de la même façon, utiliser les coefficients de Cramer au lieu des coefficients de Tschuprow.

**Illustration**

La matrice des coefficients de Tchuprow relative aux données de l'exemple 3.3 est la suivante :

$$\mathbf{T} = \begin{bmatrix} 1 & 0,08 & 0,13 & 0,18 \\ 0,08 & 1 & 0,11 & 0,16 \\ 0,13 & 0,11 & 1 & 0,12 \\ 0,18 & 0,16 & 0,12 & 1 \end{bmatrix}.$$

## 4.4 Le tableau de Burt

Le tableau de Burt est une généralisation particulière de la table de contingence dans le cas où l'on étudie simultanément $p$ variables qualitatives. Notons $X^1, \ldots, X^p$ ces variables, appelons $c_j$ le nombre de modalités de $X^j, j = 1, \ldots, p$ et posons $c = \sum_{j=1}^{p} c_j$. Le tableau de Burt est en fait une matrice carrée $c \times c$, constituée de $p^2$ sous–matrices. Chacune des $p$ sous–matrices diagonales est relative à l'une des $p$ variables ; la $j^{\text{ième}}$ d'entre elles est carrée d'ordre $c_j$, diagonale, et comporte sur la diagonale les effectifs marginaux de $X^j$. La sous–matrice figurant dans le bloc d'indice $(j, j'), j \neq j'$, est la table de contingence construite en mettant $X^j$ en lignes et $X^{j'}$ en colonnes ; le tableau de Burt est donc symétrique. Il apparaît en fait comme l'analogue qualitatif du tableau des nuages.

**Illustration**

Toujours avec les données de l'exemple 3.3, nous avons déterminé le tableau de Burt pour les trois variables série de bac, âge au bac, et durée du DEUG.