Master
Écoles d'ingénieurs

Lise Bellanger
Arthur Coulon
Richard Tomassone

Exploration de données et méthodes statistiques

Data analysis & Data mining
Avec le logiciel R

2e édition

ellipses

Références sciences

Master
Écoles
d'ingénieurs

Exploration de données et méthodes statistiques

Data analysis & Data mining
Avec le logiciel R

2e édition

Lise Bellanger
Arthur Coulon
Richard Tomassone

ellipses

Collection Références sciences

dirigée par Paul de Laboulaye
paul.delaboulaye@editions-ellipses.fr

Retrouvez tous les livres de la collection et des extraits sur www.editions-ellipses.fr

ISBN 9782340-113909
Dépôt légal : mars 2026

www.editions-ellipses.fr

AVANT-PROPOS

Cet ouvrage est le fruit de notre expérience d'enseignants de statistique appliquée et de notre collaboration à des travaux de recherche dans des domaines d'application les plus variés. À l'origine, nous avions envisagé d'y associer le terme **data mining**, alors largement utilisé. Mais ce choix nous a semblé réducteur : il ne rendait pas compte de l'ensemble des méthodes statistiques, des réflexions et des domaines d'application que nous souhaitions présenter dans cet ouvrage. L'expression **exploration des données** est apparue plus adéquate pour traduire cette ambition.

Si l'on met à part la **statistique mathématique**, la statistique n'a de sens et d'intérêt véritable que lorsqu'elle est appliquée. C'est ce que pensaient les « grands » statisticiens de la première moitié du XX^{e} siècle. A leur suite, John Tukey, statisticien un peu iconoclaste, disait que la statistique oscille toujours entre la **statistique exploratoire** qui fournit une idée qualitative des sujets analysés par une exploration minutieuse des données disponibles et la **statistique confirmatoire** qui, à partir de la précédente, met en œuvre des techniques permettant de confirmer (ou d'infirmer) des hypothèses de travail. Il disait aussi qu'il ne voyait guère de différence entre l'analyse de variance (archétype de la statistique) et l'analyse des données, à une époque où les deux approches n'étaient jamais enseignées simultanément.

Un autre élément nous paraît essentiel : placer l'objectif de l'étude au centre de toute stratégie d'analyse d'un corpus de données, avant même l'application d'une méthode particulière. En effet, rien ne permet d'affirmer qu'une seule méthode suffise à résoudre le problème posé ; c'est souvent en combinant des approches différentes et complémentaires que l'utilisateur parvient à une analyse pertinente. Il arrive même que la meilleure réponse vienne de techniques extérieures à la culture statistique classique, par exemple celles issues de l'Intelligence Artificielle. Dire que l'objectif de l'étude est essentiel, c'est aussi rappeler l'humilité nécessaire dans la pratique statistique – même pour celui ou celle qui aurait mis au point une nouvelle méthode !

En définitive, l'héritage de la pensée statistique classique tient en deux points essentiels : obtenir une estimation correcte des paramètres d'intérêt (qu'il s'agisse du temps qu'il fera demain, de la date de fabrication d'une céramique ou de la solvabilité d'un client de banque) et en apprécier la précision de la meilleure façon possible. La statistique ne fournit jamais une valeur exacte, mais seulement un domaine plausible pour l'estimation : **un chiffre dépourvu de marge d'erreur constitue une information incomplète, donc potentiellement erronée**. Les conclusions d'une étude n'ont de chances d'être valables que si le corpus analysé représente fidèlement la réalité que l'on souhaite étudier. Et si, de plus, elles sont représentatives d'un contexte plus large, elles peuvent être mobilisées dans d'autres situations.

L'ouvrage ne couvre pas toutes les facettes de la statistique appliquée, mais cherche à offrir une vision personnelle et pratique du traitement des données. Certaines approches ont été volontairement laissées de côté et ne seront qu'évoquées en conclusion. La raison en est simple : nous avons dû faire des choix puisqu'il n'était pas possible dans ce cadre, de fournir une vision exhaustive de l'ensemble des méthodes statistiques

disponibles. Pour celles utilisées plus rarement, des références sont proposées afin que le lecteur puisse s'y reporter en cas de besoin.

Le lecteur auquel s'adresse cet ouvrage doit posséder des notions de base en statistique : savoir ce qu'est une estimation, comprendre la notion d'hypothèse statistique et les tests qui y sont associés. Une pratique de l'algèbre linéaire est également requise. Le niveau attendu correspond à celui d'un Master ou d'une école d'ingénieur. Néanmoins, un lecteur moins familier du formalisme mathématique peut se concentrer sur les applications et revenir ultérieurement sur les étapes de la logique statistique présentées de manière plus formelle.

Nous supposerons également que le lecteur est familier du logiciel statistique libre ***R*** utilisé systématiquement dans l'ouvrage. Une formulation mathématique et des instructions ***R*** seront fréquemment introduites en parallèle, afin de montrer le passage de la présentation formelle à sa mise en œuvre numérique. Le choix de ce langage est lié à sa simplicité d'apprentissage, à sa très large diffusion, à sa souplesse, à la richesse de son écosystème, à sa large diffusion dans le monde académique et professionnel et à la vitalité de sa communauté. Une documentation abondante est disponible, notamment sur le site officiel du projet R (http://www.r-project.org), où l'onglet *Manuals* permet d'accéder à de nombreux documents en anglais, mais aussi en français et dans d'autres langues. Parmi les ressources francophones utiles pour apprendre à programmer avec R ou se familiariser avec sa syntaxe, on peut citer les ouvrages de Paradis (2002), Barnier (2008), Goulet (2012) et Hervé (2012). D'autres ouvrages fournissent aussi quelques rudiments de syntaxe permettant à tout utilisateur de faire rapidement ses propres programmes, par exemple Verzani (2005), Dell'Omodarme (2012), Cornillon et al. (2018), Dalgaard (2008), Robert et Casella (2011), Williams (2011) ou Chambers (2008). Venables et Ripley (2002) reste également une référence, bien qu'il concerne le langage ***S***, très proche de ***R***.

Des ouvrages récents, accessibles en ligne, illustrent également la diversité des domaines d'application de ***R*** dans la statistique moderne. Wickham et al. (2023, https://r4ds.hadley.nz/) présentent une nouvelle édition de ***R*** *for Data Science*, qui introduit une approche structurée et reproductible de l'analyse des données à l'aide du *tidyverse*. Kuhn et Silge (2023, https://www.tmwr.org/) détaillent le cadre *tidymodels*, désormais incontournable pour la modélisation prédictive. Hyndman et Athanasopoulos (2021, https://otexts.com/fpp3/) proposent un panorama complet de la prévision statistique dans un environnement ***R*** moderne, et James et al. (2021, https://www.statlearning.com/) offrent une introduction rigoureuse et pédagogique à l'apprentissage statistique. Ces références montrent que ***R*** reste un outil central, à la fois pour l'enseignement, la recherche et les applications.

De nombreux exemples seront traités tout au cours de l'ouvrage. Les fichiers de données peuvent difficilement y être inclus. Ils sont disponibles, avec une description succincte, sur le site suivante : https://R-explo-donnees.gitlab.io/website/

Enfin, dernière remarque : bien souvent la statistique appliquée est un travail commun associant une personne travaillant dans un domaine particulier et un statisticien ; il est évident que le demandeur n'a qu'une connaissance limitée des pratiques statistiques, sans quoi il ne ferait sans doute pas appel à un statisticien. Néanmoins,

c'est le demandeur qui devra prendre la décision correspondant à l'objectif de son étude : ce n'est pas le statisticien qui décide. Si un nouveau médicament doit être introduit, le statisticien peut quantifier les risques liés à son emploi, donner une mesure de ses effets secondaires ; il peut aussi voir s'il est meilleur que le précédent. Mais c'est une autorité sanitaire qui prendra la décision politique de le mettre sur le marché.

On trouvera dans l'ouvrage un certains nombre d'illustrations. D'abord des portraits, reproduction ou photographie, de certains des auteurs que nous avons cités ; généralement, si on connaît le nom d'un auteur, il est rare qu'on lui associe un visage. Puisque, pour les besoins de la composition, certaines pages seraient restées blanches, nous en avons profité pour les combler. Ensuite, quelques montages de photographies ou d'images ont été associés à une partie des exemples que nous traitons, que ce soient des céramiques, des chenilles ou du fromage, et même une vieille illustration de la Bête du Gévaudan. Cette variété n'est pas là uniquement pour distraire le lecteur habitué au formalisme mathématique, elle doit lui permettre de mesurer la diversité des champs d'application que nous présentons.

La première édition de cet ouvrage fut le fruit d'un long travail collectif. Elle n'aurait pas vu le jour sans l'aide de nombreuses personnes, que nous tenons à remercier pour la relecture de l'édition de 2014. Tout d'abord, Pierre Cazes et Roberte Tomassone, qui, chacun dans leur domaine, ont échenillé le projet : signalant des indices mal placés, corrigeant des erreurs ou améliorant la clarté des raisonnements. D'autres collègues ont relu certains passages plus proches de leurs centres d'intérêt : Didier Dacunha-Castelle, Anne Philippe, Monique Pontier et Véronique Sébille. Nos remerciements vont également à Philippe Husi, pour avoir fourni, en tant qu'archéologue, un terrain d'application et un soutien tout au long de la rédaction ; à Sylvie Karila, pour son analyse critique de la démarche ; et à Corinne Scheid, pour son travail de DAO.

Depuis, le temps a passé, et certaines des personnes qui ont accompagné ce projet avec rigueur et générosité nous ont malheureusement quittés : Pierre Cazes, Roberte Tomassone ; mais aussi Richard Tomassone, co-auteur de la première édition. Cette nouvelle édition leur est dédiée, en témoignage de notre reconnaissance et de notre souvenir ému. Elle marque également une étape importante dans la vie de l'ouvrage, avec l'arrivée d'un nouvel auteur, Arthur Coulon, qui a contribué à cette révision en profondeur tout en s'inscrivant dans la continuité du projet initial.

Cette nouvelle édition est donc une version révisée et actualisée de la précédente. Réalisée avec l'outil Quarto sous ***R***, elle associe texte et code dans une démarche de transparence et de reproductibilité des analyses. Les scripts ***R*** ont été entièrement revus et un site internet dédié accompagne désormais l'ouvrage. Plusieurs chapitres ont été enrichis, notamment le chapitre 14, qui a été complété pour prendre en compte les évolutions méthodologiques en exploration de données.

Cette nouvelle édition se veut ainsi modernisée et plus interactive, tout en demeurant fidèle à l'esprit et aux objectifs initiaux de l'ouvrage. Ce livre doit toujours autant aux étudiants curieux et passionnés, dont l'enthousiasme a non seulement été une motivation essentielle pour sa première réalisation, mais continue de nourrir cette réédition. Bien sûr, nous assumons pleinement tout ce qui aurait pu passer au travers de tous ces regards exigeants.

NOTATIONS

$\mathbf{X}$ ou $\mathbf{X}_{n\times p}$	:	matrice, éventuellement dimensions précisées par n : nombre de lignes, p : nombre de colonnes $\mathbf{X} = \mathbf{X}_{n\times p} = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ \vdots & \square & \vdots & \square & \vdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^p \\ \vdots & \square & \vdots & \square & \vdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_i^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} = \left[\boldsymbol{x}^1 \cdots \boldsymbol{x}^j \cdots \boldsymbol{x}^p\right]$
x_i^j	:	terme général de la matrice $\mathbf{X}_{n\times p}$; observation n°i $(i = 1, \cdots, n)$ de la j ème variable $(j = 1, \cdots, n)$
$\mathbf{X}^{\mathrm{T}}$	:	matrice transposée de $\mathbf{X}$
$\boldsymbol{x}_i$	:	vecteur-ligne mis en colonne associé à l'observation i ; $\boldsymbol{x}_i = [x_i^1, x_i^2, \cdots, x_i^p]^{\mathrm{T}}$
$\boldsymbol{x}^j$	:	vecteur-colonne associé à la variable j ; $\boldsymbol{x}^j = [x_1^j, x_2^j, \cdots, x_n^j]^{\mathrm{T}}$
$\mathbf{diag}(\boldsymbol{x})$	:	matrice diagonale formée des éléments du vecteur $\boldsymbol{x}$
$\mathbf{1}_p$	:	vecteur-unité d'ordre p : $\mathbf{1}_p = [1, \cdots, 1]^{\mathrm{T}}$
$\mathbf{1}_{p\times q}$	:	matrice formée de 1 à p lignes et q colonnes, $\mathbf{1}_{p\times 1} = \mathbf{1}_p$
$\mathbf{I}_n$	:	matrice identité de dimension n; $\mathbf{I}_n = \mathbf{diag}(\mathbf{1}_n)$
$det(\mathbf{A}) = \lvert\mathbf{A}\rvert$	:	déterminant de la matrice carrée $\mathbf{A}(p \times p)$
$tr(\mathbf{A})$	:	trace de la matrice carrée $\mathbf{A}$: $tr(\mathbf{A}) = \sum_{i=1}^{p} a_i^i$
$\mathbf{A}^{-1}$	:	inverse (si elle existe) de la matrice carrée $\mathbf{A}$
$E(\boldsymbol{x})$	:	espérance de $\boldsymbol{x}$
$var(\boldsymbol{x})$	:	variance de $\boldsymbol{x}$
ddl	:	degré de liberté
esb	:	estimateur sans biais
i.i.d.	:	indépendantes et identiquement distribuées
$\mathcal{N}(\mu, \sigma^2)$	:	loi Normale d'espérance $\mu \in \mathbb{R}$ et de variances σ^2
u_α	:	quantile d'ordre α d'une loi Normale centrée-réduite $N(0, 1)$
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	:	loi multiNormale de dimension p de vecteur espérance $\boldsymbol{\mu} \in \mathbb{R}^p$ et de matrice de variances-covariances $\boldsymbol{\Sigma}(p \times p)$
$T(n)$	:	loi de Student à n *ddl*
$t_{\alpha;n}$	:	quantile d'ordre α d'une loi de Student à n *ddl*
$\chi^2(n)$	:	loi du Chi-deux à n *ddl*
$\chi^2_{\alpha;n}$	:	quantile d'ordre α d'une loi du Chi-deux à n *ddl*
$F(p, n)$	:	loi de Fisher à p et n *ddl*
$f_{\alpha;p;n}$	:	quantile d'ordre α d'une loi de Fisher à p et n *ddl*
p-valeur	:	associée à la réalisation d'un test, la p-valeur (en anglais p-value) représente la probabilité d'observer une statistique de test aussi extrême, ou plus extrême, que celle observée, sous l'hypothèse nulle H_0. Pour prendre une décision, on compare la p-valeur au seuil de significativité α. $p \leq \alpha$, rejet de H_0 ; $p > \alpha$, non rejet H_0.

Avec le logiciel *R*

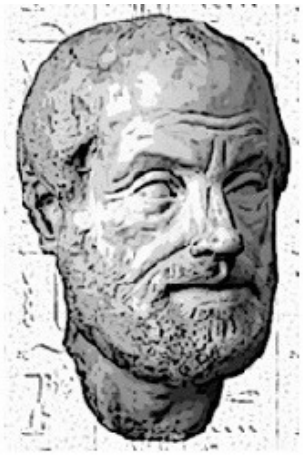
`var <- 1`	:	créer une variable `var` à laquelle on assigne la valeur 1. pareil que `var=1`. A noter que ***R*** est sensible à la casse (*case sensitive*) ce qui veux dire que les majuscule ont leur importance ("Var" est différent de "var" : `"Var"!="var"`).
`c()`	:	fonction pour créer (`c` pour *create*) un vecteur. Exmple : `Mon_vecteur=c("nom1","nom2","nom3")`
`matrix()`	:	fonction pour créer une matrice. Exmple : `matrix(c(1,2,3,4),ncol=2)` pour un matrice de taille 2×2 rempli par les valeur 1, 2, 3 et 4.
`data.frame()`	:	fonction pour créer un tableau. Exemple : `data.frame(col1=c(1,2),col2=c(3,4))`
`$`	:	dans ***R*** le symbole `$` permet d'extraire un élément d'un objet par son nom. par exemple `Tableau$Nom_colonne` extrait la colonne grâce à son nom.
`objet[i,j]`	:	extraire des valeurs grâce aux indices. Ici `objet` possède deux dimension (tableau, matrice) et `i` correspond à la ligne et `j` à la colonne. Pour un vecteur (1 dimension) on écrira `vecteur[i]`. Dans ***R*** les indices sont les numéros de ligne. Dans d'autres langages ils peuvent commencer à 0 comme avec **python**.
`;`	:	le `;` permet de signaler la fin d'une instruction. Par défaut c'est le retour à la ligne mais avec `;` on peut mettre sur la même ligne deux instructions. Dans l'ouvrage il permet de rendre les scripts plus compacts (mise en page).
`#`	:	commentaire, tout ce qui suit ne sera pas exécuté.
`library`	:	package, paquet ou bibliothèque de programmes ***R*** dévolu à des méthodes statistiques spécifiques ou à des domaines d'application particuliers. Il complète les fonctionnalités de ***R***. La fonction `library` permet de l'appeler.
`yacca::cca()`	:	fonction « `cca` » de la `library(yacca)`
`help(cca)`	:	affiche la page d'aide de la fonction `cca`. `?cca` fait la même chose.
`cca(x, y)`	:	`x` et `y`, arguments de la fonction `cca`. La liste complète des arguments et leur description se trouve dans la page d'aide de la fonction.
`sd(x)`	:	écart-type (en anglais : *standard deviation*) de l'objet `x` (un vecteur de valeurs numérique)
`dnorm`	:	`d` densité d'une loi Normale (gaussienne) ; `dnorm(x,moyenne=0, écart-type=1)`
`pnorm`	:	`p` fonction de distribution ; `pnorm(q, moyenne=0, écart-type=1)`
`qnorm`	:	`q` quantile ; `qnorm(p, moyenne=0, écart-type=1)`
`rnorm`	:	`r` génération de n variables aléatoire Normales `rnorm(n, moyenne=0, écart-type=1)`
`dchisq`	:	même usage que la loi Normale (`d`,`p`,`q`,`r`) pour la distribution χ^2
`df`		de Fisher
`dt`		de Student
`set.seed(123)`	:	En informatique, une *seed* (graine) est un nombre utilisé pour générer des séquences aléatoires. La fonction `set.seed()` de ***R*** permet de fixer cette graine, assurant ainsi la reproductibilité des résultats en produisant la même séquence à chaque exécution.

Et même à Venise, on honore la mémoire de Marino Sanuto Torsello (1260-1338), géographe et voyageur vénitien, ***initiateur de la Science Statistique*** par une plaque commémorative posée en 1883 au Ponte San Severo !

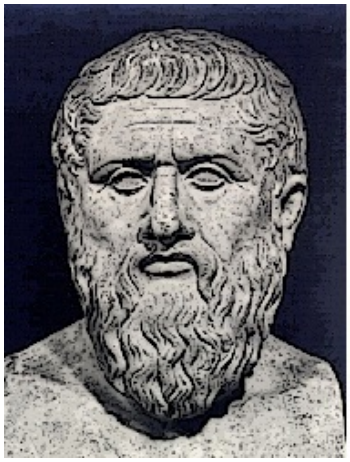
TABLE DES MATIÈRES

PARTIE I

PRÉALABLE À UN TRAITEMENT STATISTIQUE



Aristote
(-384 : -322)



Platon
(-428 : -348)

John Graunt
(1620 : 1674)

René Descartes
(1596 : 1650)

CHAPITRE 1

UNE DÉMARCHE SCIENTIFIQUE

La statistique

« Voici les chiffres communiqués par les services de la statistique et intéressant la période comprise entre le 2 juillet et le 4 septembre : 543 285 ; 6 282 826 ; 1 285 938 743 ; 601 ; 602 ; 603 ; 604 ; 605 ; 106 ; 206 ; 306 ; 406 ; 506 ; 983 ; 882 ; 780 ; 680 ; 579. Nous ne savons pas du tout à quoi se rapportent ces chiffres, mais nous sommes heureux de les communiquer à nos lecteurs qui auront ainsi toute latitude de les adapter suivant leur goût ou leur appréciation... »

Pierre DAC, *L'os à Moelle*, Juillard, Paris (1963)

Dans ce chapitre, nous allons présenter quelques idées qui vont guider notre approche de l'analyse statistique et de l'exploration des données. Après quelques réflexions sur les fondements de l'analyse statistique (§1.1), nous montrerons qu'en pratique une analyse n'est pas uniquement la réponse à une question mais qu'elle s'insère dans une démarche scientifique où il existe un aller-retour permanent entre la réflexion sur un problème et la réalité à laquelle nous sommes confrontés au sujet de ce problème : nous partons d'idées élémentaires que nous affinons progressivement (§1.2). Nous essaierons de formaliser cette démarche en montrant que cette formalisation nous oblige progressivement à nous poser des questions (§1.3). La possibilité de réaliser des analyses statistiques a été étendue à de nouveaux domaines dans lesquels la masse des informations à traiter était un écueil ; elle est désormais facilitée par l'existence de logiciels de calcul scientifique qui peuvent être facilement reliés à des logiciels de traitement de bases de données (§1.4).

1.1 Préalable à une analyse statistique

Est-ce que ce que nous voyons nous aide à mieux connaître le monde qui nous entoure ? En partie seulement, car notre vision si précise soit-elle n'est pas la **Réalité**, si tant est qu'elle existe, mais une version floue, bruitée. Pour **connaître**, il faut se poser au moins une question ; car la connaissance ne peut s'acquérir qu'en répondant à une question. La science statistique est constituée par des outils qui nous aident à mieux connaître le monde qui nous entoure en posant des questions puis en construisant des modèles traduisant celles-ci et dans lesquels nous introduisons une partie aléatoire. Comment parvenir à connaître le monde dans lequel les questions que nous posons peuvent avoir au moins une solution ? Deux approches sont possibles : doit-on partir d'une idée ou d'un fait ? Suivant ce que proposent Nisbet et al. (2009), on peut penser comme Platon ou comme Aristote : le premier croit que les choses les plus importantes dans notre existence sont au-delà de ce que nos yeux peuvent voir et que nos mains peuvent toucher ; le second croit que la connaissance d'un système complexe provient de l'étude détaillée du milieu qui l'entoure et, pour en comprendre la réalité, nous devons décomposer ce système en morceaux que nous devons décrire séparément avant de les regrouper pour comprendre le fonctionnement global. Pour Aristote, la connaissance de l'ensemble est celle de la somme de ses parties, alors que

pour Platon elle est supérieure à cette somme. La vision du premier n'est que très partiellement satisfaisante car dans un système global (comme un écosystème ou une entreprise commerciale) l'association des parties peut créer des propriétés nouvelles qui n'apparaissent que par l'existence de cette association. Techniquement ceci signifie que l'action de deux variables n'est pas toujours prévisible par la seule connaissance de chacune d'elle. C'est ce que Nisbet traduit en opposant *product-oriented knowledge-management* à *process-oriented knowledge management*. Nous en verrons des exemples tout au long de cet ouvrage.

Existe-t-il des domaines professionnels ou scientifiques desquels l'analyse statistique est absente ? Nous l'espérons. Néanmoins, ils sont plus difficiles à trouver qu'on ne le pense. Regardons autour de nous des domaines variés que rien ne semble rapprocher, sans aucun lien apparent, tous font appel à la statistique : de la physique théorique (où l'on passe du comportement microscopique de particules individuelles à un comportement macroscopique) à la datation de céramiques en archéologie, en passant par les Sciences Humaines et les Sciences de la Vie (linguistique, médecine, agronomie, écologie, etc.) et les Sciences Economiques et Sociales (l'économétrie bien sûr mais aussi la sociologie). Si nous nous aventurons vers d'autres domaines liés à l'Argent, l'intrusion est encore plus flagrante : jeux de hasard mais aussi assurance et finance et naturellement le marketing où le sondage d'opinion devient un outil pour prendre des décisions et faire des investissements. Dès lors que des informations chiffrées apparaissent et qu'il faut arriver à les maîtriser, la statistique appliquée fournit les outils techniques indispensables. Mais ne fournit-elle que des outils ? Comme d'autres l'ont dit avant nous, nous ne le pensons pas (Rao, 1994).

Analyser des données grâce à des outils statistiques est le but essentiel de la statistique appliquée. Mais pour analyser des données encore faut-il les relier à un problème que nous souhaitons étudier et à un ensemble de questions que nous nous posons au sujet de ce problème. Il faut aussi que ces questions ne soient ni vagues ni trop générales. Le statisticien travaille rarement en solitaire, il est la plupart du temps associé à une autre personne qui lui soumet un problème. Le statisticien doit donc aider celle-ci à préciser ses questions de manière à ce qu'il puisse y répondre. Pour y parvenir nous verrons qu'il faut suivre un chemin rigoureux. Ce chemin doit permettre de fournir une réponse, peut-être même plusieurs. Cette réponse est ce que l'on pourrait appeler la destination de l'analyse. Or, si arriver à destination est essentiel, bien souvent c'est le chemin suivi qui nous apporte le plus d'informations importantes en améliorant notre connaissance. C'est dans ce sens que la statistique appliquée n'est pas qu'un outil mais qu'elle peut devenir la base d'une démarche scientifique, donc rigoureuse par nature, du traitement de données avec un hasard toujours présent qui est une composante importante qu'il faut savoir apprivoiser.

Toute analyse est confrontée à trois obstacles :

1. Dispose-t-on d'un corpus de données suffisamment important ?
2. Ce corpus représente-t-il correctement l'univers à explorer ?
3. Ce corpus est-il pertinent pour répondre au problème posé ?

Si on a su franchir ces obstacles, il est alors possible de définir une stratégie pour résoudre ce problème. Les auteurs anglo-saxons parlent souvent de *process of data mining*. Nous allons étudier de façon plus détaillée cette stratégie : ce n'est pas une approche « linéaire » du problème, du type une question suivie d'une réponse. Partant de la question initiale, c'est plutôt un chemin le long duquel on s'arrête pour faire des vérifications, des *checkpoints*, éventuellement pour formuler différemment la question initiale. On avance donc de manière itérative vers une solution, si possible la meilleure, à la question posée.

1.2 Schéma d'une démarche scientifique

Dans toute démarche scientifique, nous pouvons distinguer trois phases :

- **Phase 1 :** pour atteindre le **but** $\mathcal{B}$ d'une étude, il est indispensable de partir d'un **bilan des connaissances** $\mathcal{C}$ déjà acquises au moins partiellement. Le but $\mathcal{B}$ peut être l'étude de la croissance de bactéries dans un produit alimentaire. Un **objectif** scientifique $\mathcal{O}$ peut alors être dégagé ; s'il doit être précis il n'a pas encore à se référer à une méthode particulière. L'objectif $\mathcal{O}$ peut être de prévoir le nombre de bactéries au bout d'un certain temps. Par contre, la définition de $\mathcal{O}$ implique que nous connaissions les grandes classes de questions auxquelles nous sommes capables de répondre. Une liste non limitative de questions possibles est : décrire des observations, expliquer des relations, estimer des paramètres, comparer deux ou plusieurs stratégies, classer des observations dans des groupes connus, classifier des observations, prédire le futur à partir d'observations passées, etc.
- **Phase 2 :** l'objectif étant fixé, il est possible de faire l'inventaire des modèles connus susceptibles de l'atteindre. Par **modèle**, nous ne faisons pas obligatoirement référence à une présentation formelle et abstraite (Legay, 1997). Dire qu'un échantillon est représenté par sa moyenne peut déjà être considéré comme un modèle, simpliste certes, mais un modèle. L'un de ces modèles $\mathcal{M}$ devient le premier candidat. Son choix correspond bien souvent à notre propre expérience et à notre compétence ; mais aussi à nos capacités à le traiter et à l'appliquer. Ces dernières supposent que nous puissions faire les calculs qu'il demande et que des données $\mathcal{D}$ soient disponibles pour le calibrer. Si elles ne le sont pas, il faudra les obtenir en prenant des mesures, en faisant des enquêtes, des expérimentations, etc. Mais bâtir un modèle $\mathcal{M}$ et obtenir des données $\mathcal{D}$ va nous enfermer dans un cadre relativement rigide et restreint :

  — $\mathcal{M}$ ne sera valable que sous certaines conditions que nous appellerons des suppositions [1] ;
  — L'ensemble des données $\mathcal{D}$ ne sera utilisable que si elles satisfont à certaines propriétés.

  Les méthodes statistiques que nous présenterons permettent d'étudier les propriétés intrinsèques de $\mathcal{M}$; confrontées aux données $\mathcal{D}$ elles nous

1. En anglais : *assumptions*. Notons que le terme d'*hypothèse* est souvent employé à la place de supposition ; dans un contexte statistique il existe une confusion possible avec la notion de test d'hypothèses, en anglais *hypothesis*.

permettent d'obtenir un nouveau modèle noté $\widehat{\mathcal{M}}$ qui s'appelle le **modèle estimé**. Par exemple, pour estimer le taux de survie de bactéries dans un aliment au cours du temps nous pouvons choisir pour $\mathcal{M}$ une droite ; les données $\mathcal{D}$ nous permettent d'obtenir l'équation précise, c'est-à-dire les valeurs numériques qui nous permettent de tracer la droite sur un graphique, c'est $\widehat{\mathcal{M}}$. Mais l'obtention de $\widehat{\mathcal{M}}$ n'a été faite qu'au prix d'un certain nombre de simplifications, par exemple le choix d'une droite *a priori*. Sont-elles acceptables ? C'est la dernière étape dite de *validation* $\mathcal{V}$ de cette phase ; elle aboutit à l'acceptation ou au refus de $\mathcal{M}$ et de sa réalisation $\widehat{\mathcal{M}}$. La démarche passe par une interrogation :

— Oui $\mathcal{M}$ est acceptable, nous pouvons continuer ;
— Non $\mathcal{M}$ n'est pas acceptable, il faut donc rebrousser chemin et modifier le modèle, peut-être aussi obtenir des données supplémentaires.

Nous remplaçons $\mathcal{M}$ par un nouveau modèle $\mathcal{M}^{(1)}$, voire de nouvelles données $\mathcal{D}^{(1)}$, et nous parcourons les mêmes étapes que pour le premier modèle. Au bout d'un certain temps, par améliorations successives, par itération, nous aboutissons à un modèle que nous finissons par accepter ; pour éviter une trop grande abondance de notations nous l'appellerons $\mathcal{M}$. On peut alors passer à la dernière phase.

- **Phase 3 :** puisque le dernier modèle a reçu un quitus, nous allons l'utiliser comme substitut à la réalité. Il peut être utilisé pour prédire des situations nouvelles qui peuvent se présenter. Il peut aussi servir à simuler des situations non encore explorées. Ici encore, une confrontation avec la réalité est indispensable, généralement en vérifiant si ce qui a été prédit est suffisamment proche de la réalité observée. Là encore, si nous sommes satisfaits des résultats nous continuons à utiliser $\mathcal{M}$, sinon nous le modifions à nouveau.

Donc en fin d'étude, nous avons une réponse $\mathcal{R}$ à la question initialement posée et surtout l'accumulation de nouvelles connaissances : $\mathcal{C}$ est remplacée par $\bigoplus \mathcal{C}$.

1.3 D'un but à une réalisation

1.3.1 Triplet fondamental

Du schéma précédent, il ressort :

- Un *schéma itératif* de l'analyse du problème qui était posé. Nous avons développé un *processus d'apprentissage* ;
- Un triplet $\{\mathcal{O}, \mathcal{M}, \mathcal{D}\}$ qui constitue l'outil méthodologique permettant de répondre le plus correctement possible à l'objectif simplifié $\mathcal{O}$, réponse partielle au but $\mathcal{B}$.

Dans cet ouvrage, nous nous limiterons essentiellement à la seconde phase ; nous essaierons de montrer comment à un objectif $\mathcal{O}$ nous pouvons associer un modèle $\mathcal{M}$ choisi parmi d'autres. Un modèle est la vision à un instant donné d'une partie du monde réel. Ce n'est qu'une vision partielle, idéale dont on peut simplement dire qu'elle est utile pour atteindre l'objectif fixé. Chaque modèle implique l'existence de données $\mathcal{D}$; elles mêmes peuvent servir à des modèles différents. Il existe des relations

entre les trois éléments de ce triplet (Fig. 1.1). Bien sûr, tout ce schéma doit être complété par l'introduction de moyens informatiques qui rendent le traitement des données quasiment transparent pour l'utilisateur.

A un objectif $\mathcal{O}$, nous essaierons de faire correspondre au moins un modèle $\mathcal{M}$. Mais il peut en exister plusieurs ; ce sera d'abord au statisticien de choisir le plus adapté en fonction de ses compétences et des moyens dont il dispose. Il devra préciser les conditions de son utilisation, les limites de son exploitation par un utilisateur non-statisticien. Ce dernier doit intervenir car, beaucoup mieux que le statisticien, il connaît son domaine ; il connaît aussi les limites de compétence de ceux qui peuvent l'utiliser ultérieurement de manière régulière. Il faudra aussi comprendre qu'un bon choix des données $\mathcal{D}$ est essentiel : il est inutile de les accumuler en croyant qu'un ordinateur va faire découvrir par un coup de baguette magique des richesses insoupçonnées. Il sera donc utile de réfléchir à la meilleure façon de les utiliser, par exemple pas toutes en même temps. Mais il ne faut pas perdre de vue que la qualité des données $\mathcal{D}$ est un point extrêmement important. Si les données sont trop entachées d'erreurs, contiennent des effets liés aux traitements par lots [2] , des biais dus à une mauvaise planification de leur recueil, le statisticien ne pourra pas faire de miracle ! La réponse $\mathcal{R}$ à la question posée sera de piètre qualité car aucun modèle ne parviendra à s'ajuster de manière satisfaisante aux données !

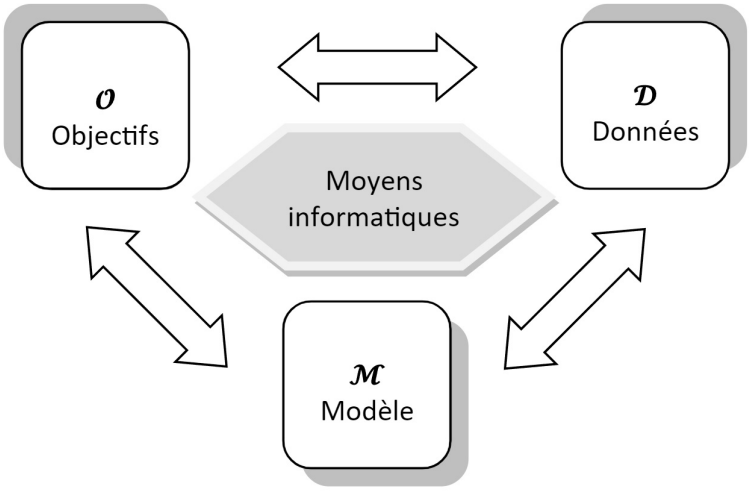


Figure 1.1 – Etapes fondamentales d'une analyse statistique.

1.3.2 D'une vision idéale à la réalité

Nous allons essayer de préciser certains éléments de la démarche que nous avons décrite sur l'exemple de croissance de bactéries, sujet important pour garantir la sécurité alimentaire des produits que nous consommons. Si $\mathcal{M}$ est vrai, nous savons (c'est notre bilan de connaissances $\mathcal{C}$ qui nous permet de le dire) que la relation est

2. Ce qu'on appelle des effets *batch* dans le traitement et l'utilisation de données de puces à ADN.

simple, de forme quasiment linéaire $\boldsymbol{y} = a + b\boldsymbol{x}$, où $\boldsymbol{x}$ est le temps (par exemple exprimé en jours) et $\boldsymbol{y}$ le logarithme du nombre de bactéries. Si $\boldsymbol{\mathcal{M}}$ était strictement linéaire, tout point (i) de coordonnées $\boldsymbol{y} = y_i$ et $\boldsymbol{x} = x_i$ serait situé sur la droite d'équation $f(\boldsymbol{x}) = a + b\boldsymbol{x}$; il suffirait alors de deux points particuliers (deux observations) pour déterminer les valeurs de a et b. Mais, ce n'est sûrement pas le cas ; si on fait une expérience pour toute une série de valeurs (à des temps fixés par l'expérimentateur) $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ auxquelles on mesure le nombre de bactéries et que leur logarithme soit $\boldsymbol{y} = \{y_1, \cdots, y_n\}$, les n points de coordonnées $\{x_i, y_i\}$ forment un nuage plus ou moins proche d'une droite si $\boldsymbol{\mathcal{M}}$ est raisonnablement correct. En règle générale, le modèle $\boldsymbol{\mathcal{M}}$ n'est pas représenté par une relation mathématique, mais par une relation plus floue. Cette relation s'obtient en introduisant dans $\boldsymbol{\mathcal{M}}$ une composante aléatoire ou bruit ou *aléa*. Plus précisément, ce que l'on observe est que $\boldsymbol{y}$ (les variables à expliquer appelée quelquefois la réponse [3]) est lié fonctionnellement à $\boldsymbol{x}$ (la variable explicative que nous appellerons plus loin régresseur [4]) par une relation plus complexe que $f(\boldsymbol{x})$ de la forme :

$$\boldsymbol{\mathcal{M}} : \ \boldsymbol{y} = f(\boldsymbol{x}, \textit{paramètres}, \textit{aléa})$$

Sous cette forme le modèle n'est pas « opérationnel », il n'est pas encore utilisable pour faire une prévision ; c'est un « instrument de discours ». Il fait apparaître les éléments importants de la *phase 2* décrite plus haut :

- y : la valeur observée (le logarithme du nombre de bactéries), image déformée d'une valeur hypothétique ;
- la fonction f qui relie $\boldsymbol{y}$ à :
  - une quantité $\boldsymbol{x}$, valeur d'une variable qui définit la **condition expérimentale** pour laquelle la valeur a été observée ;
  - des **paramètres** pour l'instant inconnus intervenant dans la fonction ;
  - l'**aléa**.

Nous noterons l'ensemble des p paramètres (ici $p = 2$) sous la forme d'un vecteur $\boldsymbol{\Theta} = (\Theta_1, \cdots, \Theta_p)^T = (a, b)^T$ et les n aléas (un pour chaque observation) par $\boldsymbol{E} = (\varepsilon_1, \cdots, \varepsilon_n)$. Nous pouvons donc réécrire formellement le modèle :

$$\boldsymbol{\mathcal{M}} : \ \boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{E})$$

Jusqu'ici rien n'est encore opérationnel ; pour avancer nous pouvons nous demander comment intervient l'aléa ; il est souvent commode de supposer que son effet est **additif**, et qu'alors $\boldsymbol{\mathcal{M}}$ peut s'écrire :

$$\boldsymbol{\mathcal{M}}^+ : \ \boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\Theta}) + \boldsymbol{E}$$

3. Il est possible d'avoir plusieurs variables à expliquer comme lors d'une *MANOVA*, dans ce cas $\boldsymbol{y}$ devient $\mathbf{Y}$.
4. Plus généralement il peut y avoir plusieurs régresseurs. Dans ce cas $\boldsymbol{x}$ devient $\mathbf{X}$.

C'est une supposition qui peut s'avérer forte ; car il pourrait tout aussi bien avoir un effet **multiplicatif** :

$$\boldsymbol{\mathcal{M}}^* : \ \boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\Theta}) * \boldsymbol{E}$$

Le choix entre $\boldsymbol{\mathcal{M}}^+$ et $\boldsymbol{\mathcal{M}}^*$ est un élément important dans la démarche que nous faisons. Il se fait soit par les connaissances *a priori* que nous avons, c'est le résultat du bilan de connaissances $\boldsymbol{\mathcal{C}}$, soit après nous être rendu compte dans le processus itératif de l'analyse que l'effet ne pouvait être d'une nature plutôt que d'une autre. Nous allons supposer que nous avons fait le choix de $\boldsymbol{\mathcal{M}}^+$; le même raisonnement pourrait se faire si le choix avait été $\boldsymbol{\mathcal{M}}^*$.

Donc si $\boldsymbol{\mathcal{M}}^+$ est vrai, pour tout couple de valeurs $\{x_i, y_i\}$, nous avons :

$$y_i = a + bx_i + \varepsilon_i, i = 1, \cdots, n$$

Si le modèle était plus complexe nous l'écririons $y_i = f(x_i, \boldsymbol{\Theta}) + \varepsilon_i$ mais ceci ne modifierait pas la suite de l'exposé. Ce qui est important est qu'un modèle additif peut toujours s'écrire :

$$\textit{Observation} = \textit{Valeur modélisée} + \textit{Aléa}$$

Mais comme "*Valeur modélisée*" et "*Aléa*" [5] qui permettent de restituer l'observation sont néanmoins inconnus, nous devons encore faire une série de suppositions sur ces deux termes :

- sur la "*Valeur modélisée*", il faut choisir une forme analytique ; c'est ce que nous avons fait plus haut en prenant une droite. C'est ce qu'on appelle la *partie contrôlée* ou encore la *partie analytique* du modèle ;
- sur l'"*Aléa*", c'est une variable aléatoire définie par une distribution de probabilité.

Il faut maintenant maîtriser l'aléa pour pouvoir l'utiliser. La statistique mathématique nous propose deux possibilités :

- choisir une forme analytique donnée et alors nous faisons de la *statistique paramétrique* ;
- donner des propriétés très générales et alors faire de la *statistique non paramétrique*.

Si nous choisissons le cadre de la statistique paramétrique, on impose généralement aux aléas ε_i des conditions du type :

- l'ensemble des valeurs modélisées passe au milieu des valeurs observées, ce qui se traduit techniquement en « l'**espérance** des ε_i » est nulle, et qui s'écrit $E(\varepsilon_i) = 0$;

5. Les physiciens emploient les termes équivalents de *signal* et *bruit*.

- la variabilité des aléas est la même dans tout le champ expérimental (le domaine de variation des x_i) ; ceci signifie que nous pensons que la précision est la même dans tout ce domaine ; ce qui s'écrit $var(\varepsilon_i) = \sigma^2 = cte$.

De surcroît, si nous supposons que toute observation apporte une information qui n'est pas déjà contenue dans une autre, les observations seront dites **indépendantes**.

Nous préciserons ces aspects tout au long de l'ouvrage ; ce qu'il faut retenir dans le cadre de la démarche générale que nous exposons c'est que toute méthode est définie dans un certain contexte. Ce contexte demande que des conditions spécifiques soient réalisées pour que la méthode choisie soit appliquée de manière réaliste.

1.3.3 Utilisation d'un modèle, sa confrontation à la réalité

Pour l'instant, nous n'avons fait que réfléchir. Cette réflexion était nécessaire, mais elle ne nous a pas encore fourni un outil opérationnel. En particulier, nous ne connaissons pas encore les valeurs des p paramètres qui constituent $\boldsymbol{\Theta}$. Nous allons devoir trouver de « bonnes valeurs » ce que l'on appelle en faire une **estimation**. Mais comment y parvenir ? Les valeurs obtenues sont-elles uniques ou en existe-t-il plusieurs ? C'est ici que la statistique mathématique, associée à l'analyse numérique, fournit des réponses. Pour que $\boldsymbol{\mathcal{M}}$ soit utile, il faut que la partie contrôlée soit la plus proche possible de la partie observée, ou que la part aléatoire soit peu importante.

Cette opération est classique en mathématique, elle consiste à minimiser une fonction des aléas ; c'est donc sur la partie $\boldsymbol{E}$ du modèle que nous devons porter notre attention. Nul besoin pour l'instant d'aller plus loin dans la présentation, disons qu'une des façons simples de le faire consiste à minimiser la somme des carrés des aléas. Cette opération numérique dont nous fournirons des détails ultérieurement nous permet d'obtenir des valeurs estimées de $\boldsymbol{\Theta}$ que nous notons $\widehat{\boldsymbol{\Theta}}$; ce qui peut s'écrire :

$$\widehat{\boldsymbol{\Theta}} = g(\mathbf{X}, \boldsymbol{y})$$

Dans cette expression $\mathbf{X}$ représente l'ensemble des valeurs (tous les x_i) où des observations des variables explicatives ont été faites et $\boldsymbol{y}$ l'ensemble des observations de la variable à expliquer (tous les y_i). Elle signifie que les valeurs des paramètres dépendent aussi bien des valeurs observées que de l'endroit où elles l'ont été ; $\mathbf{X}$ est donc le **domaine expérimental** qui en statistique classique s'appelle le plan d'expérience. Comme par définition $\boldsymbol{y}$ est constitué par des variables aléatoires, le vecteur des paramètres $\widehat{\boldsymbol{\Theta}}$ l'est aussi. Nous en avons l'estimation dans la formule ci-dessus en utilisant les données disponibles. Nous pouvons aussi en avoir d'autres caractéristiques, variances et covariances, termes que nous expliciterons plus longuement dans le prochain chapitre mais que, pour l'instant, nous mettrons sous le vocable de **précision** :

$$var\{\widehat{\boldsymbol{\Theta}}\} = \sigma^2 * h(\mathbf{X}, \boldsymbol{y})$$

La précision dépend de celle des aléas (pour l'instant inconnue), du domaine expérimental et des observations. Ce formalisme permet de comprendre que le domaine expérimental est important et même essentiel pour la qualité de l'analyse. Si les

observations d'un corpus sont mal situées dans ce domaine, elles peuvent fournir de piètres résultats pour obtenir la valeur des paramètres et de leur précision. En effet, nous pouvons maintenant estimer l'ensemble $\boldsymbol{y}$ des valeurs y_i de chaque observation par la valeur la plus probable ; son **espérance** ou **moyenne attendue** $\widehat{\boldsymbol{y}}$. Ces quantités $\widehat{y}_i$ sont obtenues à partir des valeurs de $\widehat{\boldsymbol{\Theta}}$. La confrontation entre $\boldsymbol{y}$ et $\widehat{\boldsymbol{y}}$ nous permet de voir comment le modèle, à travers ses paramètres $\widehat{\boldsymbol{\Theta}}$, nous restitue les observations ; cette confrontation nous fournit aussi une mesure de la précision de ces restitutions. Nous verrons que la comparaison entre $\boldsymbol{y}$ et $\widehat{\boldsymbol{y}}$ nous fournit une estimation $\widehat{\sigma}^2$ de σ^2, la précision inconnue [6]. Mais, mieux encore, le modèle nous permet d'explorer des situations nouvelles : au sein du domaine expérimental initial il peut nous fournir des valeurs qui auraient été observées en des « points » où aucune mesure n'a encore été faite. Dans de rares cas, il permet aussi de s'aventurer hors du domaine, on dit que l'on fait une extrapolation ; mais cette pratique est dangereuse et elle ne donne pas de protection à son utilisateur : les situations extrêmes sont toujours difficiles à modéliser pensons aux catastrophes naturelles ou aux krachs boursiers pour nous en convaincre.

Prenons un exemple médical : les observations $\boldsymbol{y}$ sont définies comme les classes d'une maladie, l'affectation à une classe entraînant des soins différents comme la nécessité ou non d'une opération ; $\mathbf{X}$ représente les résultats d'analyses cliniques que l'on fait pour décider ou non de l'opération. Pour tout nouveau patient, $\widehat{\boldsymbol{y}}$ est le résultat de la décision médicale suggérée par le modèle. Savoir si on opère ou non, connaître la précision c'est-à-dire l'incertitude du choix médical, telles sont les conséquences de l'application du modèle.

1.3.4 Sélection des variables explicatives d'un modèle

Nous avons parlé de l'importance d'un bon choix des observations, une question identique se pose pour les variables. Le choix initial ne dépend en général pas du statisticien, mais d'experts du domaine exploré. Il fait intervenir des critères comme la facilité d'obtention, la stabilité de la mesure, son coût ainsi que l'intuition du spécialiste quant au résultat attendu pour ne citer que les plus courants. Ce choix initial réalisé, le statisticien [7] peut intervenir dans la sélection des régresseurs $\mathbf{X}$ à conserver dans un modèle $\boldsymbol{\mathcal{M}}$. En fait la sélection fait intervenir ce qu'on appelle la pertinence des variables, à savoir qu'une variable est dite pertinente si son absence entraîne une dégradation importante de la qualité des résultats. Il existe trois grandes classes de méthodes de sélection (Chouaib, 2011) :

- par évaluation des caractéristiques des variables avant même de construire le modèle $\boldsymbol{\mathcal{M}}$, c'est ce qu'on appelle une **sélection par filtrage** [8]. Pour savoir si une variable est ou non pertinente, l'utilisateur étudie sa dispersion, sa relation plus ou moins importante avec d'autres variables. Mais ce choix se fait sans savoir s'il peut avoir une influence sur la qualité de $\boldsymbol{\mathcal{M}}$.
- par analyse de la qualité de l'ajustement du modèle $\boldsymbol{\mathcal{M}}$, c'est ce qu'on appelle des **méthodes enveloppantes** ; nous en parlerons abondamment au chapitre

6. Adrien-Marie Legendre publie en 1805 un essai sur la méthode des moindres carrés qui permet de comparer un ensemble de données à un modèle mathématique.
7. En fait, les statisticiens sont souvent associés, sinon remplacés, par des informaticiens qui ont investi ce domaine aux marges de leur propre discipline.
8. En anglais : *filter*.

8 sur la régression, les méthodes de ce type vont de la recherche exhaustive aux procédures pas à pas ; en fait c'est la méthode la plus courante en statistique.
- par évaluation de la pertinence d'un sous-ensemble de variables dans la construction même de $\mathcal{M}$ auquel elles sont intrinsèquement liées. Nous verrons que ces méthodes permettent une sélection interne au modèle au chapitre 9 (méthodes *Lasso*) et au chapitre 13 sur les arbres binaires. Ce sont des **méthodes** dites **intégrées** ou **embarquées**[9].

1.3.5 Choix d'un modèle et validation de résultats

Nous employons ici le terme modèle dans un sens très large, même pour des situations où y faire référence est presqu'inutile comme pour une simple analyse descriptive. Choisir un modèle est un mélange :

- de connaissances pratiques du problème étudié. Par exemple s'il existe des lois physiques connues sur la relation entre $\boldsymbol{y}$ et $\mathbf{X}$, il est bon d'en tenir compte. Dans un phénomène de croissance, on sait que celle-ci est bornée donc le choix d'un modèle linéaire, sauf pour décrire une partie du phénomène, n'est sûrement pas adéquat ;
- de connaissances théoriques sur des modèles plausibles décrivant la réalité même de façon partielle ;
- de décisions expérimentales permettant de scruter l'ensemble du phénomène. Ainsi, si la variation de $\boldsymbol{y}$ en fonction de $\mathbf{X}$ est bornée (si elle admet une asymptote de valeur inconnue y_∞, comme c'est le cas pour étudier un phénomène de croissance de bactéries) il faudra expérimenter dans un domaine dans lequel la valeur maximale est presque atteinte. Si l'expérimentateur ne le fait pas, qu'il ne puisse pas atteindre cette zone ou qu'il l'oublie, alors il ne pourra pas estimer correctement la valeur de y_∞ ;
- d'une pratique de la modélisation qui, par certains aspects, peut s'apparenter à de l'art.

Doit-on en conclure que la modélisation est un travail sans issue ? Si celui qui y recourt ignore l'existence de toutes ces difficultés, nous serions tenté de répondre par l'affirmative. Par contre, s'il n'a pas trop d'illusions, alors elle peut lui être d'un grand secours. D'autant plus que la modélisation statistique va lui fournir de nouveaux outils pour mieux assurer ses connaissances.

Des remarques précédentes, peut-être un peu désabusées, il ressort qu'il est indispensable de faire une **validation** des résultats de l'analyse. Si celle-ci porte uniquement sur un corpus, la validation n'est pas essentielle. Mais si elle déborde de son cadre réduit, il semble naturel de s'assurer de la validité des premiers résultats pour pouvoir les appliquer. Par exemple, si une banque analyse un fichier de clients pour étudier finement quels sont les critères qui lui permettront d'attribuer des prêts avec le moins de risque, il faudrait être d'une grande naïveté pour ne pas imaginer qu'une telle analyse va lui servir à sélectionner les futurs demandeurs de prêt. Dans ce cas, il est important que les résultats puissent être utilisés dans un cadre plus grand. La première

9. En anglais : *embedded*.

idée consiste à faire une nouvelle étude, indépendante de la première, pour vérifier si les premiers résultats sont confirmés. Mais ce peut être long et les statisticiens ont imaginé des méthodes de validation de modèle n'employant que la première analyse, c'est ce qu'on appelle une **autovalidation**. Nous en parlerons abondamment tout au long de cet ouvrage ; pour l'instant il suffit de savoir que trois approches sont possibles et réalisables grâce à l'émergence de nouvelles générations d'ordinateurs :

- la **validation croisée** [10] permet d'estimer la fiabilité d'un modèle en utilisant l'échantillon de taille n de plusieurs façons. La plus élémentaire consiste à le diviser en deux : la première $\mathbb{E}_1$, de taille n_1, sert à construire le modèle ; c'est un **échantillon d'apprentissage**. La seconde $\mathbb{E}_2$, de taille $n_2 (n_2 = n - n_1)$, sert à valider le modèle ; c'est un **échantillon de test** [11]. Sans connaissance particulière, le choix des observations appartenant à $\mathbb{E}_1$ et à $\mathbb{E}_2$ se fait au hasard (par tirage au sort) en tenant compte éventuellement de la structure de la population échantillonnée (existence de facteurs connus *a priori* comme le sexe, l'âge, etc.). Il est facile d'estimer l'erreur en calculant les écarts entre ce qui est observé et ce qui est estimé. Une façon un peu plus sophistiquée consiste à refaire cette opération k fois, et à prendre la moyenne des erreurs [12]. Enfin une troisième façon consiste à prendre $k = n$; elle est identique à l'approche suivante [13].
- le *jackknife* consiste à faire la même opération que précédemment mais en prenant $n_2 = 1$, et à recommencer l'opération n fois pour toutes les observations, successivement omises pour estimer les paramètres.
- Le *bootstrap* consiste à considérer que l'échantillon de taille n constitue une population dans laquelle on va de nouveau échantillonner un grand nombre de fois (disons $\mathcal{C}$ fois, avec $\mathcal{C}$ égal à plusieurs centaines) mais en ayant la possibilité de reprendre plusieurs fois la même observation, c'est un *échantillonnage avec remise*.

Jackknife et *bootstrap* sont des méthodes de **ré-échantillonnage** reposant sur l'utilisation des sous-ensembles de données disponibles ou sur un tirage aléatoire avec remise à partir de l'ensemble de données initial. Elles permettent d'effectuer de l'inférence statistique mais aussi de l'autovalidation, alors que la validation croisée n'a qu'un objectif plus limité. Il ne faut pas oublier que le *jackknife*, méthode apparue en premier, a d'abord été considéré comme une méthode permettant de réduire les biais et de construire des intervalles de confiance (Quenouille, 1956 ; Tukey, 1958). Le *bootstrap* est typiquement une méthode moderne d'inférence statistique qui n'utilise que l'information de l'échantillon et les possibilités actuelles des ordinateurs.

Jackknife et *bootstrap* font appel à des méthodes de **calcul intensif**. Elles demandent énormément de calculs pour estimer les paramètres d'un modèle (n fois pour le *jackknife*, un grand nombre de fois parmi 2^n échantillons possibles pour le *bootstrap*) : elles ne sont devenues réalisables que parce que nous pouvons nous servir presque sans limite d'ordinateurs rapides. Si ces techniques sont largement

10. En anglais : *cross validation*.
11. En anglais : *holdout method*.
12. En anglais : *k-fold cross-validation*.
13. En anglais : *leave-one-out*, noté *LOO* dans les logiciels.

répandues dans les milieux scientifiques, elles le sont beaucoup moins dans le domaine des utilisateurs.

1.4 L'obtention d'un corpus de données utilisable

1.4.1 La croissance des corpus de données

L'emploi d'outils de plus en plus élaborés implique l'existence de masse importante de données. Le volume des données ne constitue pas un but en soi ; nous avons déjà affirmé que mieux vaut les choisir avec discernement que les accumuler en espérant en extraire une information jusqu'alors invisible : « *mieux vaut un corpus bien fait qu'un corpus bien gros* ». Néanmoins, l'objectif $\mathcal{O}$ que l'on a tendance à se fixer a pour ambition d'analyser des phénomènes complexes, donc généralement décrits par un grand nombre de variables.

Certes les grands corpus de données ne sont pas d'origine récente ; des recensements agricoles ont été faits en Chine vingt trois siècles avant notre ère et des recensements de populations en Egypte cinq siècles avant. Mais ce n'est qu'au XVII^{e} siècle qu'on commence à vouloir analyser les données pour en rechercher des caractéristiques communes. Dès que des structures sociales sont créées, on collecte des données que ce soit pour décrire des phénomènes ou pour prévoir le développement d'activités. En Europe, les statistiques utilisant des masses de données sont nées soit à l'initiative de guildes marchandes soit de services d'Etat [14]. En Angleterre, dès 1662 un riche mercier londonien John Graunt analysa la mortalité dans sa ville et essaya de prévoir les apparitions de la peste bubonique sur des bases statistiques. C'est au fond un des premiers exemples récents de construction de modèles à partir d'un corpus important de données.

De nos jours, si une entreprise commerciale a pour objectif ($\mathcal{O}$) de comprendre quels sont les facteurs qui influencent son chiffre d'affaires, elle essaie d'amasser toutes les informations dont elle dispose pour analyser les diverses sources d'influence. Ces informations sont nombreuses : si elle vend de multiples produits, elle veut savoir pourquoi certains points de vente fournissent de meilleurs résultats que d'autres. Pour ce faire, elle va essayer de mettre en évidence quelles sont les variables importantes et sur quels paramètres elle doit agir pour améliorer ses résultats. Ce faisant, elle va avoir tendance à utiliser toutes les informations dont elle dispose ; mais, et c'est là que le problème se complique, ces informations ne sont généralement pas sous une forme directement accessible. Il va falloir les rechercher dans des stocks de données éparpillés en des lieux différents, sous de multiples présentations. Généralement l'ensemble de ces données existe, mais il n'a pas été conçu pour cet objectif. C'est là que des moyens informatiques importants avec des logiciels de gestion de données performants vont jouer un rôle fondamental. Vitesse certes, stockage de fichiers volumineux mais aussi outils de gestion d'un nouveau type. C'est depuis les années 1940 aux Etats-Unis et 1960 en Europe que l'informatique a permis de traiter un grand nombre de données et de croiser des séries.

14. Le nom de statistique est relativement récent, son origine remonte au XVIII^{e} siècle et semble provenir du mot allemand *Staatskunde*.

L'extraction de connaissances à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques a donné naissance à tout un vocabulaire nouveau : l'exploration de données que l'on connait aussi sous l'expression de fouille de données, forage de données, prospection de données ou enfin *data mining* ; on parle encore d'extraction de connaissances à partir de données ou ECD [15]. Si l'expression *data mining* a eu une connotation négative voire péjorative, au début des années 1960, exprimant un certain mépris des statisticiens pour les démarches de recherche de corrélation sans hypothèses de départ, vers la fin des années 1980 les choses ont évolué. C'est Rakesh Agrawal (Agrawal et al., 1993) qui a développé des concepts fondamentaux dans ce domaine ; il est surtout connu pour être un des pionniers dans le domaine important de la confidentialité des données [16] (Agrawal & Srikant, 2000). Le logiciel d'IBM dénommé *Intelligent Miner* s'est nourri de ses travaux et dans le document *IBM DB2 Intelligent Miner Tutorials* de novembre 2004 il est indiqué de manière assez grandiloquente que c'est un logiciel dont les objectifs sont : « *pour un dirigeant un moyen novateur d'avoir des idées nouvelles et précieuses sur les informations contenues dans ses bases de données, d'identifier des niches de marché pour pouvoir prendre des décisions éclairées dans le monde des affaires. C'est une manière révolutionnaire de tirer parti de l'information existant déjà dans une entreprise pour planifier une stratégie pour l'avenir* ». Piatetsky-Shapiro a largement diffusé le nom de KDD qui est alors apparu comme un concept nouveau jusqu'aux développements les plus récents (Fayyad & Piatetsky-Shapiro, 2011). Ce qui caractérise toutes ces approches est la masse souvent informe des informations à manipuler, nous sommes *at the Age of Big Data* [17] ! Il faut savoir que ces techniques, auxquelles les nouveaux outils d'apprentissage automatique [18] viennent s'ajouter, permettent à une entreprise comme **Amazon** de proposer à ses clients des produits susceptibles de les intéresser.

L'idée que nous pouvons acquérir des connaissances à partir de la « mise en données » d'un très grand nombre d'aspects de la vie (ce que la littérature américaine appelle *datafication*) a continué de se diffuser dans tous les milieux, professionnels comme scientifiques, bien au-delà des analyses pionnières de Mayer-Schönberger et Cukier (2014) [19]. Les corpus de données connaissent depuis une croissance exponentielle : qu'il s'agisse de génomique, d'astronomie, de climatologie ou encore des traces numériques produites quotidiennement par nos activités en ligne, les volumes collectés se comptent désormais en pétaoctets, voire en exaoctets. Ces données représentent le carburant de l'âge de l'information dans lequel nous sommes pleinement entrés, un matériau de base que nous pouvons traiter en quasi-temps réel grâce à des infrastructures massivement parallèles et accessibles. Nous pouvons ainsi transformer l'immense quantité d'informations qu'elles recèlent en connaissances exploitables et valorisables. Il ne faut pas oublier que, au début des années 2000, il avait fallu plus de dix ans pour décoder le premier génome humain, alors qu'aujourd'hui le séquençage complet d'un

15. En anglais : *Knowledge Discovery in Databases* ou *KDD*.
16. La question de la confidentialité est importante dans la mesure où nombre d'utilisations font appel à des traitements sur des personnes physiques (les clients d'une entreprise). En France, des règles d'utilisation ont été définies ; on les trouve sur le site www.cnil.fr de la Commission Nationale de l'Informatique et des Libertés (CNIL).
17. En français *Big Data* se traduit par *Données de Masse*.
18. En anglais : *Machine Learning*.
19. Voir aussi www.big-data-book.com.

génome se fait en une journée à coût modéré ! Mais cette révolution peut apporter le meilleur comme le pire. Le meilleur, si l'on parvient à mieux soigner certaines maladies, à prévenir des catastrophes climatiques ou à détecter suffisamment tôt un tsunami dévastateur. Le pire, si la confidentialité de nos données n'est pas préservée, au moins partiellement. En effet, notre profil numérique, somme de toutes les données nous concernant, peut nous faire interdire un prêt bancaire, nous faire refuser une opération délicate, voire alimenter des dispositifs de surveillance prédictive (Lum & Isaac, 2016 ; Mohler et al., 2011). Nous évoquerons cette question dans le dernier chapitre.

1.4.2 Les systèmes de gestion de base de données

On accumule les informations de taille importante dans des bases de données. Pour les stocker et les partager entre plusieurs utilisateurs, il faut des *systèmes de gestion de base de données* ou **SGBD**. Ces systèmes permettent d'avoir une qualité des informations dont ils assurent la pérennité et la confidentialité. Un *SGBD* permet d'effectuer un certain nombre d'opérations élémentaires, rarement complexes en elles-mêmes, mais caractérisées par la masse énorme des données à traiter, comme le fichier des clients d'une entreprise de vente par correspondance. Mais si le **stockage** des informations constitue la première opération, il faut ensuite pouvoir les retrouver, éventuellement les modifier, voire les trier ou les transformer. Il existe un langage spécifique **SQL** [20] pour effectuer des opérations sur des bases de données. C'est d'abord un langage de **manipulation** permettant de rechercher, d'ajouter, de modifier ou de supprimer des données dans les bases de données. Il permet aussi de **modifier l'organisation** des données. Il permet aussi d'effectuer ce qu'on appelle des **transactions** : par exemple, une transaction peut être une réservation ou un achat. La base de données passe au cours de cette opération d'un état initial à un autre état ; la suite des opérations doit obéir à certaines contraintes qui, si elles ne sont pas satisfaites, aboutissent à l'échec de la réservation ou de l'achat. Naturellement, l'état de la base doit être permanent : un incident technique ne doit pas pouvoir entraîner l'annulation des opérations effectuées durant la transaction. Enfin, il doit assurer le contrôle des données en autorisant ou en refusant leur accès à certains utilisateurs. *SQL* a été créé en 1974, puis normalisé depuis 1986, il est reconnu par la grande majorité des *SGBD*. Enfin, au-delà d'assurer la cohérence et la confidentialité des informations, d'éviter de les perdre, il doit pouvoir les transmettre à d'autres logiciels, par exemple d'analyse statistique. Il doit comporter au moins une **interface graphique** [21] pour faciliter le dialogue homme-machine ; selon sa nature le *SGBD* peut aussi contenir des langages de programmation beaucoup plus élaborés.

Parmi les applications utilisant des *SGBD* on peut citer : les guichets automatiques bancaires, les logiciels de réservation d'*Air France* ou de la *Fenice* de Venise, ou encore les services Web. Il existe un grand nombre de systèmes allant du logiciel payant au libre, cloud ou *serverless*, avec ou sans SQL. Un des plus utilisés est MySQL (Dubois et al., 2004), en concurrence avec Oracle [22] et PostgreSQL.

20. Pour *Structured Query Language*.
21. En anglais : *GUI* pour *Graphical User Interface*.
22. *Oracle Corporation* est une entreprise américaine créée en 1977, rachetée en 2010 par *Sun Microsystems* pour la modique somme de 7.4 milliards de $. Cette dernière avait auparavant racheté *MySQL AB* qui diffusait MySQL. Elle détient la moitié du marché mondial.

L'objet de notre travail n'est pas dans cette direction, mais nous tenons à insister sur le fait que la préparation de corpus, pouvant être ultérieurement correctement traités, passe souvent par cette phase de « *data management* » qu'un statisticien a tendance à juger avec condescendance ou pour le moins à trouver très fastidieuse. C'est pourtant celle qui prend le plus de temps, le calcul scientifique ultérieur étant généralement beaucoup plus rapide.

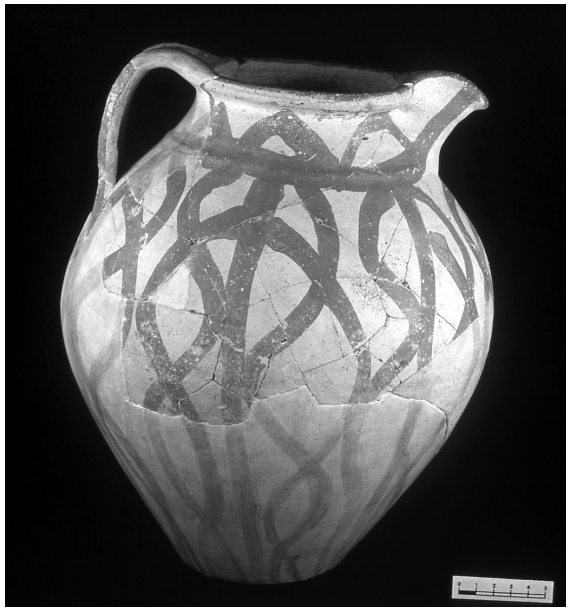
1.5 Organisation de l'ouvrage

Cet ouvrage est essentiellement consacré à ce qui s'appelait l'analyse des données multidimensionnelles. Il recouvre plusieurs aspects selon la structure des données de base. Ces données de base, nous le verrons et surtout le préciserons au chapitre 2, sont décrites par des observations sur lesquelles l'utilisateur relève des caractéristiques. Dans le domaine médical, sur un patient (ce sera ici l'observation élémentaire) on relève non seulement des caractéristiques civiles mais aussi des analyses médicales (ce sera ici des variables). Cette présentation simplifiée induit que l'ensemble des méthodes s'appliqueront à des tableaux à deux entrées (horizontale/verticale ou ligne/colonne). Le plan est donc le suivant :

- Les chapitres 2 et 3 sont consacrés à la présentation de quelques outils indispensables avant toute analyse ; nous évoquerons des questions essentielles sur le choix des variables (faut-il ou non les transformer ?), sur la possibilité qu'elles soient erronées, sur le fait que certaines puissent être partiellement manquantes ;
- Les chapitres 4, 5 et 6 sont consacrés aux méthodes d'analyse factorielle dans lesquelles les lignes pourraient être permutées de même que les colonnes sans que les résultats d'analyse en soient modifiés ; nous dirons que le tableau de données n'est **pas structuré** ;
- Le chapitre 7 est consacré au même type de tableau (toujours non structuré) à la recherche de séparation entre lignes (ou entre colonnes) pour faire des regroupements, c'est-à-dire des classes, c'est le chapitre sur la classification ;
- Les chapitres 8, 9 et 10 sont consacrés aux tableaux dans lesquels les colonnes ne jouent plus des rôles symétriques (le tableau est **structuré en colonnes** c'est le champ important des régressions et de l'analyse canonique) ;
- Les chapitres 11 et 12 prennent l'optique symétrique où les lignes sont dans des groupes connus (il existe des malades et des non malades, de bons payeurs et mauvais payeurs) c'est le domaine de la discrimination et du classement où le tableau est **structuré en lignes** ;
- Au chapitre 13, nous présenterons les arbres binaires qui sont une manière différente de combiner régression et classement.
- Enfin au chapitre 14, nous rappellerons quelques points fondamentaux de notre démarche, nous évoquerons quelques champs d'application que n'avons pas abordés, et quelques perspectives de développement qui nous semblent importantes en statistique appliquée dans un proche avenir.

L'organisation générale est schématisée sur la figure 1.2.

Bien sûr nous n'avons pas visé à l'exhaustivité du domaine de l'analyse des données et nous n'avons pas couvert tous les champs potentiellement utiles. Le lecteur pourra trouver d'autres ouvrages tels que celui de Kendall et al. (1983) de loin le plus complet, ou celui de Saporta (2011) qui, en français, est suffisant pour nombre de personnes. Pour avoir une vision plus moderne, celui de Hastie et al. (2009) est sans doute la référence à consulter. Ceux qui voudraient approfondir les relations de la statistique avec la philosophie des sciences peuvent consulter, outre les textes classiques comme Good (1988), des travaux plus récents tels que Gelman et Hennig (2017).



Céramique et tessons (Philippe Husi, UMR 7324 CITERES-LAT, 2013).

Fichier : `Ceram18`

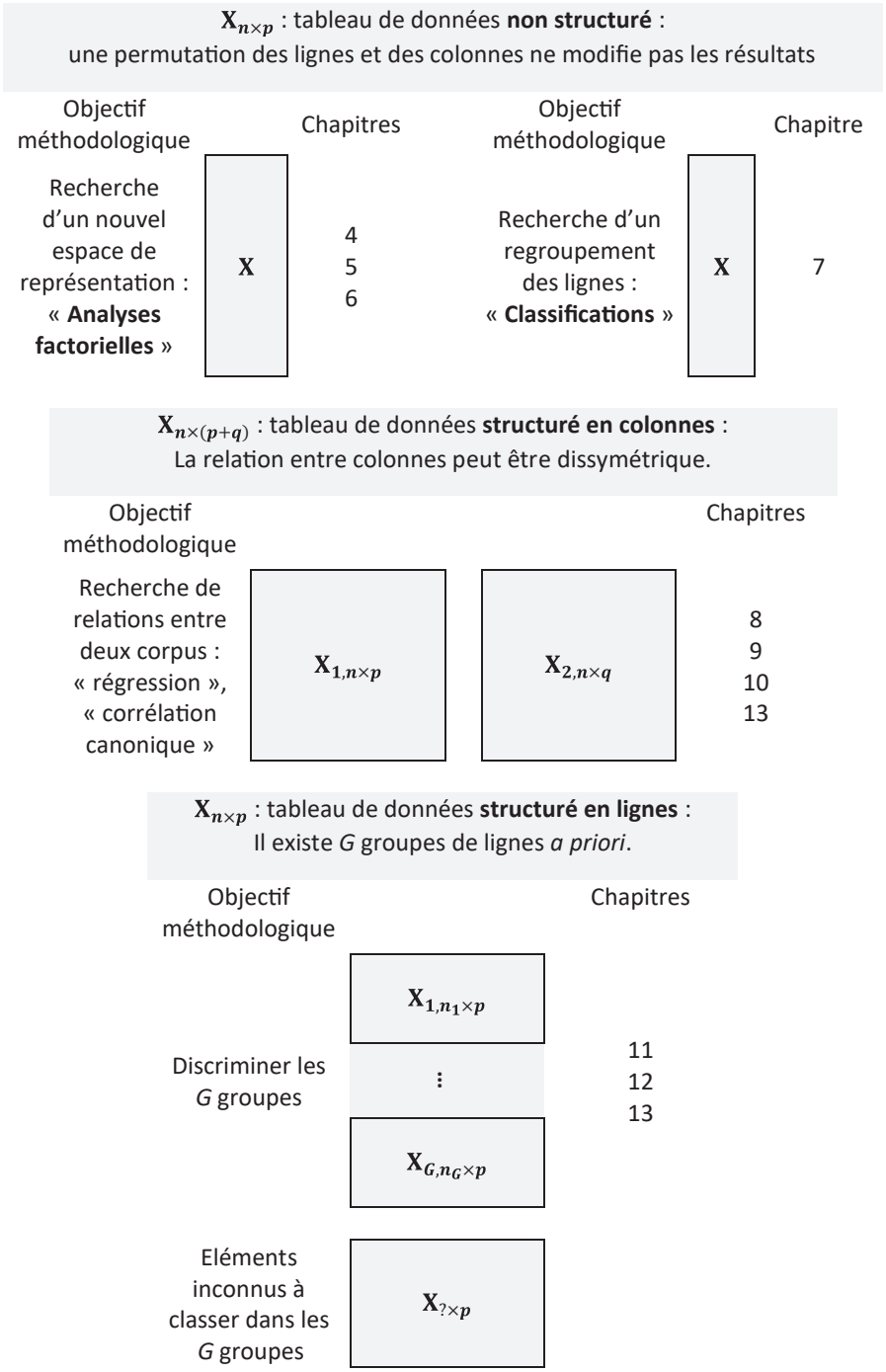


FIGURE 1.2 – Organisation de l'ouvrage en fonction de la structure du tableau de données.

Conclusion : la modélisation statistique est une forme d'art qui s'appuie sur un certain nombre d'outils mathématiques. C'est un instrument indispensable pour faire un apprentissage d'une réalité toujours complexe. Par sa capacité à simplifier la description d'une « situation », elle permet d'en déceler les traits essentiels. Il n'existe pas d'outil « toujours meilleur » pour n'importe quel utilisateur. Les outils les plus utiles sont ceux qui rendent aisé le plus grand nombre de tâches dans une exploration de données qu'un utilisateur doit réaliser.

Ce qu'il faut retenir :

- L'*observation* est la combinaison d'une valeur contrôlée inconnue ou signal et d'un aléa ; ce n'est pas la valeur modélisée du signal.
- L'*aléa*, défini par une distribution de probabilité, est un élément essentiel de toute modélisation.
- L'*estimateur* d'un paramètre a des propriétés liées à la distribution de l'aléa.
- Un corpus de données (une expérience) n'est qu'une *réalisation* particulière, avec un autre corpus les valeurs sont différentes.
- La confrontation d'un modèle formel et d'une expérience fournit une *estimation* ; une autre expérience en fournit une autre.
- La *validation* est une phase indispensable dans toute modélisation.
- La *préparation* d'un corpus utilisable pour une analyse de données est la phase la plus longue et la plus délicate.
- La qualité des résultats est fonction de la qualité des données à l'entrée ou *garbage in, garbage out*.

CHAPITRE 2

LES OUTILS DE REPRÉSENTATION D'UN ÉCHANTILLON

Lorsque l'on fait des statistiques, on travaille en général sur des **observations** sur lesquelles sont mesurées ou observées des **variables** en plus ou moins grand nombre. Une observation est l'unité élémentaire ; nous verrons qu'elle peut avoir des formes multiples : une personne ou un groupe de personnes, un animal ou une portée d'animaux, une plante, un organe ou une seule cellule. Une variable est une caractéristique ou une propriété qu'il est possible de mesurer. Mesurer quelque chose signifie que nous faisons un modèle numérique de la chose à mesurer : nous assignons un nombre à un niveau de la caractéristique à mesurer. Le poids d'une personne est une variable ; nous assignons à deux personnes de même poids la même valeur numérique.

Dans ce chapitre nous allons définir ce qu'est la structure d'un ensemble **observations / variables** que nous appellerons tableau de données (§2.1) ; nous verrons ensuite comment résumer un tableau de données à l'aide d'un petit nombre de caractéristiques soit unidimensionnelles (§2.2) soit multidimensionnelles (§2.3) ; nous présenterons un outil commode pour représenter le tableau de données (§2.4), puis nous dresserons un bilan des aspects importants de ce chapitre (§2.5).

2.1 Structure d'un tableau de données

2.1.1 Questions préalables

Toute expérience ou toute enquête se fonde sur la mesure de quantités, qui sont les réalisations d'une ou de plusieurs variables. Mais avant toute analyse de ces données, il faut se poser un certain nombre de questions préalables :

- **Que sont les données ?** Sont-elles en nombre important, quelle est leur précision ?
- **Quelles sont les unités ?** Kilomètre ou mile marin, degré Celsius ou degré Farenheit, etc. ?
- **D'où proviennent-elles ?** Les valeurs sont-elles raisonnables ? Des valeurs comme 0.7 ou 44 pour le poids d'une dinde en kilogrammes indiquent-elles une simple erreur de transcription ou mettent-elles en doute l'ensemble des relevés ? Pour un relevé pluviométrique, l'emplacement du pluviomètre peut-il entraîner des variations importantes des mesures obtenues ?
- **Comment et quand ont-elles été mesurées ?** Des biais sont-ils possibles ? L'observateur sait-il arrondir les nombres ? Que signifie la température d'un jour donné, est-ce la valeur maximale, la valeur à une heure précise ? Les enregistrements sont-ils différents en fin de semaine, ont-ils été réalisés par la même personne ?

- **Comment ont-elles été acquises ?** A-t-on mesuré toutes les unités de la population, ou bien a-t-on fait un échantillonnage ? Dans ce dernier cas, le choix de l'unité expérimentale peut-il avoir une influence sur le résultat ?
- **Existe-t-il une structure logique entre les observations ?** S'agit-il du poids d'un animal mesuré à plusieurs époques ou du poids d'animaux différents ? Les unités ont-elles un rapport entre elles, comme les animaux d'une même portée ou les fruits d'une même variété ? Les observations ont-elles en commun certains facteurs d'environnement ?

C'est seulement quand on a répondu à toutes ces questions, que l'on peut pousser plus loin l'analyse.

2.1.2 La forme du tableau de données

Il est généralement toujours possible de présenter des données sous la forme d'un tableau rectangulaire à I lignes et J colonnes ; quelquefois nous utiliserons les notations n au lieu de I et p au lieu de J. Formellement, nous pouvons écrire qu'un tableau de données $\mathbf{X}$ est de la forme :

$$\mathbf{X} = [x_i^j], i = 1, \cdots, I \ (ou\ n)\,; j = 1, \cdots, J \ (ou\ p)$$

Ci-dessus, les deux indices i et j repèrent la valeur x_i^j qui se trouve à la ligne i et à la colonne j. Mais la représentation d'une ligne ou d'une colonne est différente :

- une ligne i, notée $\boldsymbol{x}_i = [x_i^1 \ \cdots \ x_i^j \ \cdots \ x_i^J)]^{\mathrm{T}}$, représente une unité de base : c'est l'**observation élémentaire** sur laquelle ont été mesurées, observées ou contrôlées J caractéristiques.
- une colonne j, notée $\boldsymbol{x}^j = [x_1^j \ \cdots \ x_i^j \ \cdots \ x_I^j)]^{\mathrm{T}}$, représente la valeur soit d'une **variable**, soit d'un **indicateur** [1]. La différence entre les deux provient de l'origine expérimentale de l'observation, c'est-à-dire de la façon dont on l'a obtenue : relever l'âge d'une personne fournit une variable, mais sélectionner des personnes d'un âge fixé, fournit un indicateur. C'est un problème que nous évoquerons plusieurs fois dans cet ouvrage, dans la mesure où la distinction peut ne pas être toujours évidente.

2.1.3 Notions de type

La valeur d'une variable, comme celle d'un indicateur, dépend de la structure mathématique qui permet de la définir. Est-elle définie sur l'ensemble des nombres réels $\mathbb{R}$ ou simplement sur l'ensemble des réels positifs $\mathbb{R}^+$? Ne peut-elle prendre qu'un nombre limité de valeurs entières ? Quelquefois, la valeur ne correspond pas à une structure numérique : on peut toutefois s'y ramener par un simple **codage**. Par exemple, pour une couleur qui peut ne présenter que les trois caractéristiques vert, jaune ou marron, on peut soit établir la correspondance suivante :

$$\{vert, jaune, marron\} \Longleftrightarrow \{1, 2, 3\}$$

soit utiliser pour chaque réalisation une variable à deux valeurs seulement, 0 ou 1. Chaque modalité de la variable définit une variable dichotomique ; la correspondance est alors :

1. Dans la terminologie du modèle linéaire, on emploie le mot **facteur**.

Tableau 2.1 – Exemple de passage d'un codage simple à un codage disjonctif complet.

	codage initial	vert	jaune	marron
vert :	1	1	0	0
jaune :	2	0	1	0
marron :	3	0	0	1

Pour un traitement statistique, on utilise ce dernier codage appelé **codage disjonctif complet**. Ainsi un objet vert sera analysé par l'ensemble des trois valeurs $\{1, 0, 0\}$, un jaune par $\{0, 1, 0\}$ et un marron par $\{0, 0, 1\}$, comme indiqué dans le tableau 2.1.

Physiquement, une mesure numérique sur $\mathbb{R}$ ne peut prendre qu'un nombre fini de valeurs. En pratique, ce qui distingue les valeurs sur $\mathbb{R}$ de celles sur $\mathbb{N}$ dépend essentiellement du nombre de valeurs possibles : si ce nombre est assez grand (> 10), on pourra considérer que la mesure est définie sur $\mathbb{R}$, sinon elle l'est sur $\mathbb{N}$; le nombre d'enfants d'une famille est généralement défini sur $\mathbb{N}$ alors que la taille de la portée d'un animal prolifique peut l'être sur $\mathbb{R}$. Néanmoins, rien ne s'oppose à dire que le nombre moyen d'enfants d'une famille européenne est de 1.8, bien que cette valeur n'ait aucun sens pour une famille particulière. La limite entre la définition sur $\mathbb{R}$ ou sur $\mathbb{N}$ dépend du contexte de l'application qui devra toujours être précisé. Cette remarque ne s'applique pas à un codage ; elle n'a de sens que pour une caractéristique sur laquelle on peut faire une opération arithmétique comme la sommation, donc un calcul de moyenne. Pour une utilisation statistique, il est utile de remplacer la notion de structure mathématique par celle de type de variable, schématisée dans le tableau 2.2.

Tableau 2.2 – Les différents types de variables ou d'indicateurs, nature et exemples.

Variable ou indicateur		Type	Nature	Exemples
quantitative :	–	réel	continue	poids, taille, revenu
	–	entier	discrète	nombre d'enfants
qualitative :	–	polytomie ordonnée	ordinale	mauvais/moyen/bon, malade/atteint/sain
	–	polytomie non ordonnée	nominale	profession, couleur, variété végétale
	–	dichotomie	binaire	vrai/faux, oui/non, présence/absence

2.1.4 Représentation d'un tableau de données

Le tableau $\mathbf{X}$ est donc constitué par deux sous-tableaux :

$$\mathbf{X} = [\mathbf{X}^1 \vdots \mathbf{X}^2]$$

$\mathbf{X}^1$ est un tableau à I lignes et p colonnes ; $\mathbf{X}^2$ est un tableau à I lignes et q colonnes. Donc $\mathbf{X}$ est un tableau à I lignes et $J = p + q$ colonnes. L'unité expérimentale $n^\circ I$, la ligne i de, $\mathbf{X}$ est donc formée d'un ensemble de p variables et de q indicateurs ; les p

variables définissent le vecteur observation $\mathbf{X}_i^1$ représentant la $i^{\text{ème}}$ ligne de la matrice $\mathbf{X}^1$, lui-même éventuellement repéré par un **vecteur de contrôle** $\mathbf{X}_i^2$. L'ensemble des méthodes statistiques se rattachent :

- soit à des **méthodes unidimensionnelles** pour l'analyse séparée de chaque colonne de $\mathbf{X}^1$;
- soit à des **méthodes multidimensionnelles** [2] pour l'analyse simultanée de l'ensemble des colonnes de $\mathbf{X}^1$.

Dans les deux cas, l'existence d'un vecteur de contrôle induit une structure dans le tableau de données : connue *a priori*, elle sera généralement utilisée dans l'analyse des données ; mais elle pourra aussi ne fournir que des informations non prises en compte dans l'analyse, néanmoins utiles pour l'interprétation.

Quelquefois, on peut compléter un premier tableau $\mathbf{X}$ par d'autres observations sur les mêmes J colonnes. Si ces observations ne sont pas prises en compte dans l'analyse initiale, on dira que l'on a affaire à des **observations passives** ou **observations supplémentaires**, par opposition aux premières qui sont des **observations actives**. Naturellement, on peut aussi ajouter d'autres colonnes à $\mathbf{X}$ sur les mêmes unités de base ; on parlera de **variables passives** ou **variables supplémentaires** (les secondes) et de **variables actives** (les premières). La même chose peut se produire pour des indicateurs. Schématiquement nous pouvons écrire que le tableau $\mathbf{X}$ est complété de la manière suivante :

$$\begin{bmatrix} \mathbf{X} & \mathbf{X}^{Csup} \\ \mathbf{X}^{Lsup} & / \end{bmatrix}$$

Bien que nous ne l'ayons pas indiqué, le symbole « / » peut être remplacé par des variables supplémentaires associées à des observations supplémentaires.

2.2 Résumés unidimensionnels

Nous débuterons ce paragraphe par l'étude des variables quantitatives ; puis nous verrons ensuite de manière plus succincte au §2.2.5 le cas des variables qualitatives. Avant d'étudier un tableau comme $\mathbf{X}^1$, il est toujours indispensable d'analyser séparément chacune de ses colonnes ; pour des variables quantitatives, on commence par étudier séparément chaque **distribution** des variables du tableau, soit p distributions. Il existe plusieurs manières de représenter ces distributions. A titre d'exemple regardons le tableau 2.3 qui représente 20 valeurs de 6 variables chimiques de composition d'eau minérale. On peut étudier ce tableau de deux manières : à l'aide de graphiques ou à l'aide de résumés numériques ; nous commencerons par les premiers.

Dans ce paragraphe nous allons considérer les colonnes de $\mathbf{X}^1$ l'une après l'autre, même si ne nous interdisons pas des regards croisés ; aussi, pour éviter l'usage d'indices superflus, notons $\mathbf{x}^j$ $(j = 1, \cdots, p)$; les n valeurs de la colonne j (n est noté I au début du §2.2.4).

L'analyse des colonnes de $\mathbf{X}^2$ est aussi possible, s'il existe une structure dans le tableau de données. Elle permet de voir si cette structure induit un équilibre entre les différents groupes définis par $\mathbf{X}^2$; ainsi, lorsque $\mathbf{X}^2$ n'a qu'une colonne (alors $q = 1$), si

2. On trouve aussi l'adjectif *multivarié*, provenant directement du *multivariate* anglais.

la variable a k modalités, l'écriture disjonctive de $\mathbf{X}^2$ comportera k colonnes associées aux k variables définies comme dans le tableau 2.1. Par exemple, si c'est une couleur : existe-t-il autant d'observations dans chaque niveau de couleur ? Nous verrons plus loin que l'existence d'un équilibre est souvent un élément important pour faire les « meilleures » statistiques. Dans la suite de ce chapitre, nous n'utiliserons pas cette possibilité de structuration, nous utiliserons la notation $\mathbf{X}$ à la place de $\mathbf{X}^1$. De plus, nous noterons n le nombre lignes (=observations, au lieu de I) et p celui des colonnes (=variables).

TABLEAU 2.3 – `Eaux1` : Echantillon de 20 eaux minérales non gazeuses pour 6 variables chimiques.

Eau	HCO3	SO4	Cl	Ca	Mg	Na
Aix	341	27	3	84	23	2
Bec	263	23	9	91	5	3
Cay	287	3	5	44	24	23
Cha	298	9	23	96	6	11
Cri	200	15	8	70	2	4
Cyr	250	5	20	71	6	11
Evi	357	10	2	78	24	5
Fer	311	14	18	73	18	13
Hip	256	6	23	86	3	18
Lau	186	10	16	64	4	9
Oge	183	16	44	48	11	31
Ond	398	218	15	157	35	8
Per	348	51	31	140	4	14
Rib	168	24	8	55	5	9
Spa	110	65	5	4	1	3
Tho	332	14	8	103	16	5
Ver	196	18	6	58	6	13
Vil	59	7	6	16	2	9
Vit	402	306	15	202	36	3
Vol	64	7	8	10	6	8

2.2.1 Graphiques

Nous verrons plus loin que les méthodes statistiques dépendent beaucoup des **suppositions** que l'on peut faire sur les variables. Il est donc indispensable de faire, avant tout calcul, des graphiques permettant de voir si ces suppositions sont plus ou moins bien respectées, c'est ce qu'on appelle une analyse exploratoire des données [3]. Ces graphiques permettent de répondre aux deux questions suivantes :

- La distribution des données est-elle sensiblement Normale, c'est-à-dire symétrique et relativement concentrée autour de la position centrale ?
- Existe-t-il des **valeurs suspectes** [4] ?

3. En anglais : *Exploratory Data Analysis* (EDA).
4. En anglais : *outliers*.

On peut faire quatre types de graphiques : un **histogramme**, un **graphique de densité**, un **graphique de Normalité**[5] et un **diagramme en boîte**[6].

Un histogramme n'est parlant que si le nombre n d'observations est suffisamment élevé (> 100), aussi est-il utile de le remplacer par le graphique de densité qui en est une version lissée dont nous parlerons plus loin.

Un graphique de Normalité est un graphique des valeurs ordonnées de la variable en fonction des quantiles correspondants de la loi Normale standard (de moyenne 0 et de variance 1) : la distribution est d'autant plus proche de la Normalité que le graphique est proche d'une droite.

Un diagramme en boîte est aussi très facile à interpréter : si nous appelons $Q_{0.25}$, $Q_{0.50}$ (qui est la **médiane**) et $Q_{0.75}$ les valeurs telles que le quart, la moitié et les trois-quarts des observations leur soient inférieures (ce que nous appellerons plus loin les quartiles) on trace une boîte comme celle de la figure 2.1 ; les données sont celles de la variable `Ca` du fichier `Eaux1` (Tab. 2.3). Les extrémités indiquées par « | » sont à des positions : pour l'inférieure à $Q_{0.25} - 1.5(Q_{0.75} - Q_{0.25})$ et pour la supérieure à $Q_{0.25} + 1.5(Q_{0.75} - Q_{0.25})$ et les valeurs qui sont extérieures sont indiquées par des « ∘ ». On repère ainsi très facilement les valeurs suspectes ou extrêmes, et on voit très rapidement si la distribution est symétrique. Ici $Q_{0.25} = 53.25$; $Q_{0.50} = 72.00$ et $Q_{0.75} = 92.25$. Apparaissent à droite deux observations suspectes correspondant aux valeurs 157 et 202.

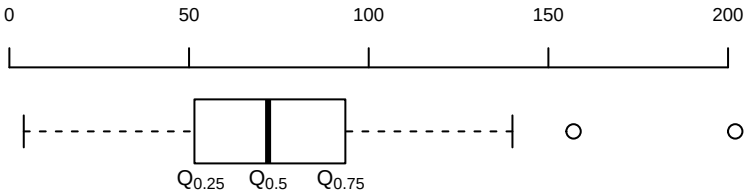


Figure 2.1 – Représentation de la boîte à moustaches pour la variable `Ca` du fichier `Eaux1`.

La comparaison de plusieurs distributions est facilitée. Ainsi si nous regardons les deux premières variables du fichier de données `Eaux1` la distribution de `HCO3` apparaît assez symétrique, alors que celle de `SO4` ne l'est pas du tout. Les quatre autres boîtes à moustaches donnent une vision plus claire de la dissymétrie (Script 2.1, Fig. 2.2).

Script 2.1 – `Eaux1` : Boîtes à moustaches.

```
## Importation des données Eaux1
Eaux1 <- read.table("data/Eaux1.txt", header=TRUE, row.names=7)
## Tracé des six boîtes à moustaches (Fig. 2.2)
boxplot(Eaux1, col = "white")
## horizontal=TRUE pour des boxplots horizontaux (comme Fig. 2.1)
# boxplot(Eaux1$Ca, col = "white", horizontal=TRUE)
```

5. Qu'on appelle aussi *qq-plot*.
6. Qu'on appelle aussi *boîte à moustaches* ou *Box and Whisker plot* ou *box-plot*.

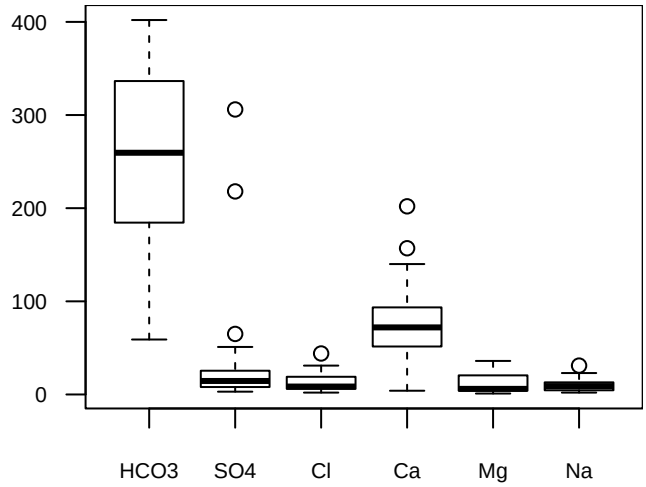
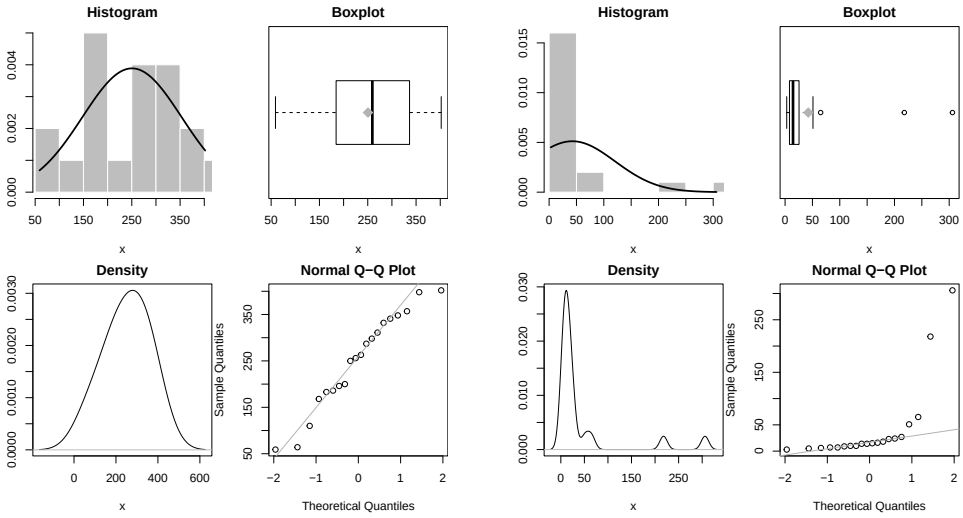


FIGURE 2.2 – `Eaux1` : Boîtes à moustaches des six variables.

Un regard sur les graphiques complets de ces mêmes variables montre bien les différences entre les deux : `HCO3` a une distribution voisine de la Normale, alors que `SO4` est très dissymétrique avec un fort étalement vers les grandes valeurs (SCRIPT 2.2, FIG. 2.3). Les graphes de Normalité vont dans le même sens : la forme de la distribution de `HCO3` est beaucoup proche de celle de la loi Normale que celle de `SO4` (FIG. 2.3).

SCRIPT 2.2 – `Eaux1` : instructions permettant de faire les graphiques des figures 2.3a et 2.3b.

```
Dist.forme <- function(x) ## Création d'une fonction 'Dist.forme'
{                         ## la fonction possède un argument : x
  old=par();par(mfrow=c(2,2)) ## configuration en 2 lignes 2 colonnes
  x.naomit <- na.omit(x)      ## supprime les valeurs manquantes
  # graphe 1 : histogramme et densité correspondante
  hist(x.naomit, col = "cyan4", border = "white", prob = TRUE,
       xlim = c(min(x.naomit),max(x.naomit)), main = "Histogram")
  curve(dnorm(x, mean=mean(x.naomit), sd = sd(x.naomit)), add=TRUE,
        lwd=2, col="indianred")  ## ajoute courbe densité en rouge
  # graphe 2 : boxplot
  boxplot(x.naomit, col="white", horizontal=TRUE, main="Boxplot")
  points(mean(x.naomit), y=1, col="orange", pch=18, cex=2) ## moyenne
  # graphe 3 : Densité
  iqd=summary(x.naomit)[5]-summary(x.naomit)[2] ## écart interquartile
  plot(density(x.naomit,width=2*iqd),xlab="x",ylab="",main="Density")
  # graphe 4 : qqplot
  qqnorm(x.naomit); qqline(x.naomit, col="indianred")
  par(mfrow=old$mfrow) ## Paramètre mfrow réinitialisé par défault
}
## Tracé des quatre graphiques réalisés sur les variables HCO3 et SO4
Dist.forme(Eaux1$HCO3); Dist.forme(Eaux1$SO4)
```



(a) HCO3 (b) SO4

FIGURE 2.3 – Eaux1 : Graphiques d'analyse exploratoire des données (*EDA*) des deux variables HCO3 (a) et SO4 (b).

2.2.2 Paramètres numériques classiques

Pour compléter ces visions graphiques, il faut calculer des paramètres numériques traduisant le mieux possible la distribution, c'est-à-dire les caractéristiques de ces n valeurs ; on peut regrouper ces caractéristiques selon trois types d'indices : **position**, **dispersion** et **forme**. Pour chaque type, plusieurs paramètres sont possibles avec des intérêts pratiques plus ou moins importants.

Paramètres de position : Ils définissent la tendance générale de la distribution. Deux paramètres principaux sont utilisés, avec des propriétés différentes : la moyenne et la médiane d'une variable quantitative $\boldsymbol{x}$. La **moyenne**, notée $\bar{x}$, est définie par :

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

Si on a n_i valeurs x_i identiques, et si le nombre total est $n = \sum n_i$, on associe à chaque observation une **pondération** $p_i = n_i/n$, et donc $\sum p_i = 1$; alors on écrit :

$$\bar{x} = \sum_{i=1}^{n} p_i x_i$$

Naturellement, quand toutes les observations sont individualisées, $p_i = 1/n$. On appelle **variable centrée** la variable $\boldsymbol{x}_{\mathrm{C}}$ de coordonnées : $x_i - \bar{x}$; la moyenne de cette variable est donc nulle.

La **médiane**, notée Me, est la valeur qui divise les n observations en deux parties égales (50 % lui sont inférieures et 50 % supérieures), elle est déterminée lorsque l'on

a rangé les n observations par ordre croissant :

$$x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[n-1]} \leq x_{[n]}$$

alors si n est impair et égal à $2m+1$, la médiane vaut :

$$Me = x_{[m+1]}$$

et si n est pair et égal à $2m$, la médiane vaut :

$$Me = (x_{[m]} + x_{[m+1]})/2$$

Cette formulation montre que le calcul de la médiane revient à pondérer les valeurs rangées de la façon suivante : toutes les valeurs ont une pondération nulle $p_{[j]} = 0$, sauf pour :

- n impair, alors $p_{[m+1]} = 1$ et
- n pair $p_{[m]} = p_{[m+1]} = 1/2$

En fait, la médiane est un paramètre beaucoup plus stable que la moyenne et beaucoup moins sensible qu'elle aux valeurs suspectes car les valeurs extrêmes n'ont pas de poids dans son calcul. On dit que c'est un paramètre **robuste** (Script 2.3).

Script 2.3 – `Eaux1` : détermination de la médiane de `HCO3`.

```
Eaux1$HCO3                  ## valeurs de HCO3 dans le fichier
# [1] 341 263 287 298 200 250 357 311 256 186 183 398 348 168 110 332 196 59 402 64
rank(Eaux1$HCO3)            ## rang des valeurs de HCO3 dans le fichier
# [1] 16 11 12 13 8 9 18 14 10 6 5 19 17 4 3 15 7 1 20  2
sort(Eaux1$HCO3)            ## valeurs rangées par ordre croissant de HCO3
# [1] 59 64 110 168 183 186 196 200 250 256 263 287 298 311 332 341 348 357 398 402
median(Eaux1$HCO3)          ## médiane de HCO3
# [1] 259.5
```

La dernière remarque conduit à imaginer d'autres paramètres en jouant de façon moins drastique sur les pondérations des valeurs rangées. On peut ainsi calculer une **moyenne équeutée**[7], en supprimant un certain nombre de valeurs extrêmes, en général un pourcentage de n, par exemple 5 % ; si nous prenons un pourcentage de 50 % nous obtenons la médiane. Voici ce que donnent ces calculs sur les deux variables HCO3 et SO4 (Script 2.4) :

Script 2.4 – `Eaux1` : moyennes et moyennes équeutées des deux variables `HCO3` et `SO4`.

```
## HCO3                                  ## SO4
mean(Eaux1$HCO3);                        mean(Eaux1$SO4)
# [1] 250.45                             # [1] 42.4
mean(Eaux1$HCO3, trim=0.5);              mean(Eaux1$SO4, trim=0.5)
# [1] 259.5                              # [1] 14.5
mean(Eaux1$HCO3, trim=0.05);             mean(Eaux1$SO4, trim=0.05)
# [1] 252.6667                           # [1] 29.94444
```

7. En anglais : *trimmed mean*.

Une autre idée est celle de la **winsorisation** [8] qui ne modifie pas la pondération, mais qui donne à un certain nombre (ou à un certain pourcentage) de valeurs extrêmes la même valeur. Par exemple, on donne à $x_{[1]}$ la même valeur que celle de $x_{[2]}$ et à $x_{[n]}$ la même valeur que celle de $x_{[n-1]}$. Ces différents paramètres ont des propriétés intéressantes et, quelquefois, ils peuvent remplacer la moyenne.

Quand les observations sont discrètes ou regroupées en classes de même amplitude, on définit le **mode** comme la valeur ou la classe de fréquence maximale. Pour une variable quantitative, c'est la valeur la plus probable. Mais contrairement aux autres paramètres de position, le mode peut ne pas être unique : on parlera alors de **distribution multimodale**. Cet inconvénient peut devenir un avantage, puisque la présence de deux valeurs de mode peut être l'indice d'un mélange provenant de deux populations différentes. Pour une distribution Normale : Moyenne = Médiane = Mode.

Une grande différence entre ces paramètres (en particulier entre moyenne et médiane) est un indice certain de non-Normalité de la distribution.

Paramètres de dispersion : Ils définissent la variabilité de la distribution. Le plus utilisé est la variance notée s^2, définie par :

$$s^2 = var(\boldsymbol{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$$

La division par $(n-1)$ sera justifiée ultérieurement pour des raisons statistiques ; quelquefois la division est faite par n, la variance est alors la moyenne des carrés des écarts à la moyenne [9]. Nous utiliserons toujours le facteur $(n-1)$, sauf cas particulier justifié par le contexte ; de toute manière, dès que n est assez grand la différence est minime. On note s l'**écart-type** qui est la racine carrée de la variance, et s'exprime donc dans la même unité que la variable étudiée (Script 2.5).

Script 2.5 – `Eaux1` : variances et écart-types des deux variables `HCO3` et `SO4`.

```
## HCO3                                  ## SO4
var(Eaux1$HCO3, na.rm=T);                var(Eaux1$SO4)
# [1] 10529.63                           # [1] 6079.516
sd(Eaux1$HCO3);                          sd(Eaux1$SO4)
# [1] 102.6140                           # [1] 77.97125
```

On appelle **variable centrée réduite** notée $\boldsymbol{x}_{\mathrm{CR}}$ la variable de coordonnées : $(x_i - \bar{x})/s$; la moyenne de cette variable est donc nulle et son écart-type vaut 1.

Un autre paramètre de dispersion est l'**étendue** notée E, définie par :

$$E = x_{[n]} - x_{[1]}$$

L'étendue est donc la différence entre la valeur maximale et la valeur minimale des x_i : c'est un paramètre très sensible aux **valeurs extrêmes** ; on peut lui substituer des paramètres dérivés de la **distribution équeutée**. En pratique, on considère

8. https://fr.wikipedia.org/wiki/Winsorisation.
9. En anglais : *standard deviatio (sd)*.

surtout des paramètres dérivés des α**-percentiles** définis par la valeur de la variable telle que la proportion α ($\alpha < 0.5$) des observations lui soit inférieure. Si on prend la valeur entière (arrondie à l'entier le plus proche) de $\nu = \alpha n$, alors l'α-percentile $Q_\alpha = x_{[\nu]}$. L'importance des queues de distribution peut être étudiée par l'écart inter α-percentile :

$$D_\alpha = Q_{(1-\alpha)} - Q_\alpha$$

On peut démontrer, sous des conditions très larges (KENDALL et al., 1983), des propriétés des distributions de ces quantités ; en particulier, pour la médiane ($\alpha = 0.5$), on obtient la variance :

$$var(Me) = \frac{\pi}{2} var(\bar{x})$$

Ce résultat nous indique que la variance de la médiane est moins précise que celle de la moyenne.

Un paramètre de dispersion, souvent utilisé et beaucoup plus stable que l'étendue, est l'**écart interquartile** :

$$D_{0.25} = Q_{0.75} - Q_{0.25}$$

C'est celui-ci qui est représenté dans le graphique de la boîte à moustaches.

Nous mettons à part un autre paramètre de mesure de la variabilité, appelé **coefficient de variation**, qui est le rapport exprimé en % de l'écart-type à la moyenne :

$$CV(\%) = 100 s/\bar{x}$$

Il peut être intéressant pour comparer deux échantillons d'une même variable ; de plus, étant sans dimension, il donne une mesure absolue de la variabilité. Néanmoins, il peut être d'un usage dangereux. En effet, la valeur de la moyenne est modifiée si on ajoute une constante à la mesure d'une variable ; ce n'est pas le cas de la valeur de l'écart-type. Donc CV sera différent selon que l'on fournit une température en degré Celsius ou en degré Fahrenheit puisqu'une température de t degrés Fahrenheit correspond à $5(t-32)/9$ degrés Celsius.

Paramètres de forme : Leur intérêt provient principalement du fait que leur valeur théorique pour la loi Normale est connue. On les utilise donc beaucoup pour vérifier le caractère gaussien d'une distribution ce sont :

- le coefficient d'asymétrie [10] ou d'*obliquité* estimé par :

$$b_1 = \frac{m_3}{{m_2}^{3/2}} = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^3\right] \Big/ \left(\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right)^{3/2}$$

  a. < 0 si mode $> \bar{x}$; donc une queue de distribution étalée vers la gauche ;
  b. $= 0$; pour une distribution symétrique ;
  c. > 0 si mode $< \bar{x}$; donc une queue de distribution étalée vers la droite.

Notons qu'une forte asymétrie dans une distribution est un indicateur d'hétérogénéité au sein de la population étudiée.

10. En anglais : *skewness*.

- le coefficient d'aplatissement [11] :

$$b_2 = \frac{m_4}{{m_2}^2} = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^4\right] \Big/ \left(\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right)^2$$

  a. < 3 si moins aplatie qu'une loi Normale ;
  b. $= 3$; si loi Normale ;
  c. > 3 si plus aplatie qu'une loi Normale.

Des valeurs respectives des coefficients d'asymétrie et d'aplatissement différentes de 0 et 3 indiquent que la distribution du caractère est plus ou moins éloignée de la distribution Normale ; évidemment la taille de l'échantillon joue un rôle si on souhaite faire un test de l'écart à la Normalité (SCRIPT 2.6).

SCRIPT 2.6 – **Eaux1** : résumé par variable et calcul des coefficients d'asymétrie et d'aplatissement.

```
## Résumé par variable : Min., 1Quart., Médiane, Moyenne, 3Quart., Max.
summary(Eaux1)

## `e1071` nécessaire pour skewness() & kurtosis()
## `dplyr` nécessaire pour summarise() & across()
library(e1071); library(dplyr)                                          ①

## écart-type : s
summarise(Eaux1,across(1:6, list(sd = sd)))
## asymétrie : b1
summarise(Eaux1,across(1:6, list(skewness = skewness)))
## aplatissement : la fonction kurtosis renvoie b2 - 3
summarise(Eaux1,across(1:6, list(kurtosis = kurtosis)))
```

① `across` permet d'étendre l'application de la fonction "à travers" les colonnes 1 à 6 (`1:6`).

L'ensemble de ces résultats pour les six variables de **Eaux1** est fourni au tableau 2.4 [12].

TABLEAU 2.4 – **Eaux1** : paramètres des 6 variables.

variable :	HCO3	SO4	Cl	Ca	Mg	Na
$\mathbf{X}_{[1]}$	59.00	3.00	2.00	4.00	1.00	2.00
$Q_{0.25}$	185.25	8.50	6.00	53.25	4.00	4.75
Me	259.50	14.50	8.50	72.00	6.00	9.00
$\bar{x}$	250.45	42.40	13.65	77.50	11.85	10.10
$Q_{0.75}$	334.25	24.75	18.50	92.25	19.25	13.00
$\mathbf{X}_{[24]}$	402.00	306.00	44.00	202.00	36.00	31.00
s	102.61	77.97	10.59	48.09	11.09	7.32
b_1	-0.33	2.45	1.24	0.76	0.93	1.24
$b_2 - 3$	-1.02	4.84	1.02	0.39	-0.55	1.14

Le coefficient d'aplatissement de la fonction `kurtosis` renvoie $(b2-3)$ et doit donc être interprété non plus en termes de valeurs supérieures, égales ou inférieures à 3 ; mais plutôt par rapport à 0.

11. En anglais : *kurtosis*.
12. Les valeurs des coefficients d'asymétrie et d'aplatissement n'ont ici qu'une signification limitée ; il faudrait davantage d'observations pour se rendre compte si l'éloignement à la loi Normale a ou n'a pas un sens.

2.2.3 De l'intérêt des transformations des variables

Tous les paramètres présentés ci-dessus permettent de définir des indices synthétiques d'une distribution. Nous verrons ultérieurement que de nombreuses méthodes d'analyse statistique, notamment les méthodes paramétriques, se basent sur l'hypothèse de Normalité de la distribution de la variable étudiée. Même si on peut bien souvent ne pas en faire un usage exclusif en ayant recours à des méthodes statistiques qui ne requièrent pas la Normalité des variables (par exemple les méthodes non paramétriques) ; il est souvent bien utile de s'en rapprocher en particulier en faisant une **transformation normalisatrice** des variables. Le type de transformation à adopter dépendra bien sûr de l'allure de la distribution de fréquences des données brutes. En particulier une distribution étalée vers les grandes valeurs, donc dissymétrique vers la droite, peut suggérer que si l'on utilisait une *transformation logarithmique*, on serait plus proche du référent Normal. Ceci conduit naturellement à choisir la moyenne des logarithmes des valeurs comme paramètre de position ; celle-ci est la **moyenne géométrique** notée g_x, des valeurs de départ :

$$ln(g_x) = \frac{1}{n}\sum_{i=1}^{n} ln(x_i) \Rightarrow g_x = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

De la même façon, si c'est l'inverse qui suit une distribution Normale, un calcul du même type fournit la **moyenne harmonique** [13] notée h_x :

$$\frac{1}{h_x} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i} \Rightarrow h_x = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

Si toutes les valeurs x_i sont positives, on peut montrer que :

$$h_x \leq g_x \leq \bar{x}$$

Il est à noter que les transformations normalisatrices courantes possèdent la propriété importante de réduire l'**hétéroscédasticité** des données, c'est-à-dire de stabiliser leur variance. Dans de nombreuses situations, la transformation normalisatrice n'opère pas une normalisation complète mais contribue à rendre la distribution de fréquences plus symétrique, ce dont on peut se satisfaire dans un certain nombre d'analyses.

Nous verrons au chapitre suivant qu'il est possible de faire un certain nombre de transformations simples avant de commencer le traitement véritable.

2.2.4 Estimation de la densité

Sur la figure 2.4, nous pouvons voir que le tracé d'un histogramme ne fournit pas de résultats très interprétables si le nombre n d'observations est trop faible. En effet pour le tracer on doit se fixer un nombre de classes, donc un intervalle de variation h dans lequel toutes les observations ont la même densité. La densité en un point est donc la valeur n_x/nh où n_x est le nombre de points de la classe. Son calcul vérifie donc la formule générale :

13. Naturellement pour ces deux moyennes $x_i > 0, \forall i$.