

Collection

Statistique
et probabilités
appliquées

Pierre Lafaye de Micheaux
Rémy Drouilhet
Benoît Liquet

Le logiciel R

Maîtriser le langage

Effectuer des analyses (bio)statistiques

> Im(y-x)

> eigen(X)\$val

> demo

> boxplot(x)

> t.test(x)

> vecn <- function (n)



Deuxième
édition

> apply (X,FUN=mean,MARGIN

Lavoisier
hermes

Le logiciel R

**Maîtriser le langage
Effectuer des analyses (bio)statistiques
2^e édition**

Pierre Lafaye de Micheaux
Rémy Drouilhet
Benoît Liquet

Le logiciel R

Maîtriser le langage
Effectuer des analyses (bio)statistiques
2^e édition

Lavoisier
hermes

editions.lavoisier.fr

Pierre Lafaye de Micheaux

Département de mathématiques et de statistique
Université de Montréal
Pavillon André-Aisenstadt
2920, chemin de la Tour
Québec H3T 1J4
Canada

Rémy Drouilhet

B.S.H.M.
1251, avenue Centrale
BP 47
38040 Grenoble Cedex 9

Benoît Liquet

School of Mathematics and Physics
The University of Queensland
St Lucia, Brisbane 4072
Australia

;E4@ , +) *žšž & (šžž #*žš

© >ŠafW 2^e édition, 2014

Imprimé en France

Cet ouvrage est soumis au copyright. Tous droits réservés, notamment la reproduction et la représentation, la traduction, la réimpression, l'exposé, la reproduction des illustrations et des tableaux, la transmission par voie d'enregistrement sonore ou visuel, la reproduction par microfilm ou tout autre moyen ainsi que la conservation des banques de données. La loi française sur le copyright du 9 septembre 1965 dans la version en vigueur n'autorise une reproduction intégrale ou partielle que dans certains cas, et en principe moyennant le paiement des droits. Toute représentation, reproduction, contrefaçon ou conservation dans une banque de données par quelque procédé que ce soit est sanctionnée par la loi pénale sur le copyright.

L'utilisation dans cet ouvrage de désignations, dénominations commerciales, marques de fabrique, etc. même sans spécification ne signifie pas que ces termes soient libres de la législation sur les marques de fabrique et la protection des marques et qu'ils puissent être utilisés par chacun.

La maison d'édition décline toute responsabilité quant à l'exactitude des indications de dosage et des modes d'emploi. Dans chaque cas il incombe à l'utilisateur de vérifier les informations données par comparaison à la littérature existante.

Maquette de couverture : Jean-François Montmarché

Détail du tableau : Bloc Images

Collection
Statistique et probabilités appliquées
dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

Comité éditorial :

Aurore Delaigle

Département de mathématiques
et de statistique
Université de Melbourne
Victoria 3010
Australie

Christian Mazza

Département de mathématiques
Université de Fribourg
Chemin du Musée 23
CH-1700 Fribourg
Suisse

Christian Genest

Département de mathématiques
et de statistique
Université McGill
Montréal H3A 2K6
Canada

Stephan Morgenthaler

École Polytechnique Fédérale
de Lausanne
Département de Mathématiques
1015 Lausanne
Suisse

Marc Hallin

Université libre de Bruxelles
Campus de la Plaine
CP 210
1050 Bruxelles
Belgique

Louis-Paul Rivest

Département de mathématiques
et de statistique
Université Laval
Québec G1V 0A6
Canada

Ludovic Lebart

Télécom-ParisTech
46, rue Barrault
75634 Paris Cedex 13
France

Gilbert Saporta

Conservatoire national
des arts et métiers
292, rue Saint-Martin
75141 Paris Cedex 3
France

Dans la même collection :

- *Statistique. La théorie et ses applications*
Michel Lejeune, avril 2004
- *Optimisation appliquée*
Yadolah Dodge, octobre 2004
- *Le choix bayésien. Principes et pratique*
Christian P. Robert, novembre 2005
- *Régression. Théorie et applications*
Pierre-André Cornillon, Éric Matzner-Løber, janvier 2007
- *Le raisonnement bayésien. Modélisation et inférence*
Éric Parent, Jacques Bernier, juillet 2007
- *Premiers pas en simulation*
Yadolah Dodge, Giuseppe Melfi, juin 2008
- *Génétique statistique*
Stephan Morgenthaler, juillet 2008
- *Maîtriser l'aléatoire. Exercices résolus de probabilités et statistique, 2^e édition*
Eva Cantoni, Philippe Huber, Elvezio Ronchetti, septembre 2009
- *Pratique du calcul bayésien*
Jean-Jacques Boreux, Éric Parent, décembre 2009
- *Statistique. La théorie et ses applications, 2^e édition*
Michel Lejeune, septembre 2010
- *Probabilités et processus stochastiques*
Yves Caumel, janvier 2011
- *Analyse statistique des risques agro-environnementaux*
David Makowski, Hervé Monod, septembre 2011
- *Statistique appliquée aux sciences de la vie*
Valentin Rousson, janvier 2013
- *Modélisation et évaluation quantitative des risques en actuariat*
Étienne Marceau, janvier 2013

À Dominique, à Luka et à Mathias
À mes parents

À tous ceux qui ont contribué, contribuent et contribueront
à éveiller nos consciences

À Pierre et à sa persévérance

Avant-propos

Cet ouvrage est fondé sur les notes d'un cours dispensé pendant quelques années à l'Institut universitaire de technologie de Grenoble 2, au sein du département Statistique et informatique décisionnelle (STID). Il a donc été « digéré » pour la première fois, dans une version très imparfaite, par les étudiants de ce département que nous remercions ici. Sans l'intérêt témoigné par ces derniers, cet ouvrage n'aurait probablement pas vu le jour. Nous voulons également vivement remercier notre collègue et ami Michel Lejeune, qui a réussi à nous convaincre de travailler à la rédaction d'un manuscrit à soumettre aux éditions Springer. Nous souhaitons aussi souligner l'importance du hasard qui a permis que les trajectoires des trois auteurs de ce livre se croisent dans un même lieu, pendant quelques années. L'expérience humaine et scientifique qui a résulté de cette rencontre a été très enrichissante, et chacun des auteurs a pu apporter des compétences complémentaires ayant permis de venir à bout du travail considérable qu'a nécessité la rédaction de cet ouvrage. Nous tenons enfin à remercier ici très chaleureusement Matthieu Dubois, un collègue et ami, chercheur en psychologie expérimentale et féru de **R** et de l'environnement Macintosh qui a été le premier à lire ce livre dans sa version quasi finalisée et nous a conseillé de nombreuses améliorations.

L'information contenue dans ce livre a été choisie et organisée de la meilleure façon possible afin d'être **exhaustive** tout en étant également **assimilable** par le lecteur. Cet ouvrage peut ainsi servir comme support d'un cours sur le logiciel **R** à un niveau de débutant à avancé. Une emphase particulière a été mise sur la forme du livre, ce qui, à notre sens, permet d'en faciliter la compréhension. Il devrait aussi pouvoir être utilisé comme un support d'auto-apprentissage par tout autodidacte. Notons que la présentation de l'ouvrage sera majoritairement indépendante de tout système d'exploitation. Toutefois, quelques chapitres seront destinés principalement à des utilisateurs du système d'exploitation Microsoft Windows. Nous pensons également utile de donner, par endroits, des compléments pour les utilisateurs de Linux ou de Macintosh.

Les chapitres du livre sont tous structurés de la même manière. Chaque chapitre débute par un petit encart indiquant les pré-requis nécessaires à la lecture dudit chapitre ainsi qu'un descriptif succinct du contenu du chapitre.

Les notions théoriques sont agrémentées de nombreux exemples et également parsemées de pauses invitant à pratiquer directement sur l'ordinateur ce qui a été vu. Chaque chapitre se termine enfin par une partie de contrôle de l'acquisition des connaissances sous la forme d'un encadré de termes à retenir, suivie d'une section d'exercices théoriques à faire sur feuille, et pouvant servir de questions à un examen sur table. Une fiche de travaux pratiques est également fournie en fin de chapitre. Celle-ci permet de vérifier que les compétences pratiques ont bien été assimilées. Notez que les exercices et les travaux pratiques doivent être traités uniquement avec les notions apprises dans les chapitres précédents.

La trame séquentielle du livre se déroule comme suit. Après une brève introduction destinée à mettre le lecteur en appétit, et la présentation de quelques jeux de données qui seront exploités tout au long de l'ouvrage pour illustrer l'utilisation de **R**, la première partie du livre est ensuite dédiée à l'apprentissage des concepts principaux du logiciel **R** : organisation des données, importation et exportation, manipulations diverses, accès à la documentation, représentations graphiques, programmation et maintenance. Cette partie consiste donc à « faire ses gammes » sur **R**.

La seconde partie du livre est consacrée à l'utilisation du logiciel **R** dans quelques contextes mathématiques et statistiques. Cette partie devrait être lue après les chapitres de la première partie, mais elle devrait tout de même se révéler accessible aux utilisateurs possédant déjà quelques notions de **R**. Elle contient les instructions **R** nécessaires pour quelques-uns des principaux cours de statistique et de mathématiques jusqu'à la licence (couvrant par exemple le programme en IUT de statistique et informatique décisionnelle en France) : calcul matriciel, intégration, optimisation, statistiques descriptives, simulations, intervalles de confiance et tests d'hypothèses, régression linéaire simple et multiple, analyse de la variance.

Notons enfin que chaque chapitre de statistique dans la seconde partie s'appuie sur un ou plusieurs jeux de données réelles, gracieusement mis à disposition par l'ISPED (Institut de santé publique, d'épidémiologie et de développement de Bordeaux) et présentés en début d'ouvrage, qui en rendent ainsi l'apprentissage plus concret et plus attractif. Nous en profitons pour remercier particulièrement toute l'équipe pédagogique du master de santé publique de l'ISPED. Ces données, ainsi que plusieurs fonctions développées spécialement pour le livre, et qui y sont présentées ou utilisées, sont disponibles dans un *package* **R** associé à l'ouvrage qui s'appelle `LeLogicielR`. Nous remercions également Mohamed El Methni et Taghi Barumandzadeh pour le matériel qu'ils nous ont fourni dans la rédaction du chapitre sur l'ANOVA.

Deuxième édition

Nous tenons à remercier Hubert Raymondaut pour nous avoir donné la motivation nécessaire pour écrire cette seconde édition, qui s'accroît de près de 200 pages. Plusieurs erreurs mineures ont été corrigées, certaines notions clarifiées et de nombreuses astuces ou renvois vers d'autres ressources ont été ajoutés au fil du texte.

La section A.4, intitulée « L'interface graphique de **R** (GUI) », a été tronquée et une nouvelle section A.5 intitulée « Mes premiers pas en **R** » a été ajoutée. Dans cette dernière, nous décrivons l'utilisation de l'outil **RCommander**, un package permettant l'utilisation de **R** via des menus, puis expliquons comment utiliser au mieux **R** via sa console.

Dans le Chapitre 2, une nouvelle section 2.4, intitulée « Lecture/écriture dans les bases de données », a été ajoutée.

Dans le Chapitre 3, la section 3.4 a été déplacée après la section 3.7. Elle devient donc la nouvelle section 3.7. Une nouvelle section 3.8, intitulée « Création de fonctions », a été ajoutée après cette section suivie d'une nouvelle section 3.9, intitulée « Représentation des nombres à virgule fixe, flottante », expliquant les problèmes numériques pouvant survenir du fait des limites de représentation des nombres sur un ordinateur. De plus, un TP sur la création de fonctions (le F-) a été ajouté à la toute fin du TP du Chapitre 3.

Dans le Chapitre 6, une nouvelle section 6.5, intitulée « Interfacer **R** et **C/C++** ou **Fortran** », fait son apparition juste avant l'ancienne section 6.5 « Gestion de son activité de développement » qui s'intitule désormais « Débogage de fonctions » et porte le numéro 6.6. Le contenu de toute cette section a été modifié et largement augmenté. L'ancienne sous-section 6.5.1 « Débogage de fonctions » de la version 1 devient la sous-section 6.6.1 « Débogage de fonctions en **R** pur ». Nous avons aussi rajouté une section 6.7 intitulée « Calcul parallèle et calculs sur cartes graphiques ».

Le titre du Chapitre 10 a été changé en « Variables aléatoires, lois et simulations : une meilleure compréhension grâce aux spécificités de **R** » pour être plus représentatif de son contenu.

Pour finir, la correction de tous les exercices et de tous les TPs a été intégrée dans l'ouvrage, ce qui en fait très probablement le manuel le plus complet à ce jour sur le logiciel **R**. Celui-ci pourra être utilisé pour former les lycéens français dans le cadre du nouveau programme national, ainsi que les étudiants des classes préparatoires et de l'université. Il permet toujours de mener ses lecteurs à un stade avancé de maîtrise du logiciel.

Parcours différenciés

Nous avons mentionné explicitement, à l'aide du symbole †, les sections plus délicates ou moins fondamentales pouvant être écartées lors d'une première lecture de l'ouvrage, sans pour autant nuire à la compréhension et à la maîtrise du logiciel **R**.

Notez que ce livre a d'abord été pensé pour être lu par des étudiants issus de formations mathématiques ou statistiques. Toutefois, nous proposons ci-dessous, pour les étudiants ou les chercheurs ayant suivi un parcours plus « appliqué », d'adopter un parcours différencié pour le cœur de l'ouvrage. La lecture des sections délicates sera également omise.

PARTIE I : LES BASES DU LOGICIEL

- a) Les concepts de base, l'organisation des données (chapitre 1).
- b) Importation-exportation et production de données (chapitre 2).
- c) Manipulation de données (chapitre 3).
- d) **R** et sa documentation (chapitre 4).
- e) Techniques pour tracer des courbes et des graphiques (chapitre 5).
- f) Maintenance des sessions (chapitre 7).

PARTIE II : STATISTIQUES ÉLÉMENTAIRES

- a) Variables aléatoires, lois et simulations (chapitre 10).
- b) Statistique descriptive (chapitre 9).
- c) Intervalles de confiance et tests d'hypothèses (chapitre 11).
- d) Régression linéaire simple et multiple (chapitre 12).
- e) Analyse de variance élémentaire (chapitre 13).

PARTIE III : CONCEPTS AVANCÉS

- a) Mathématiques de base : calcul matriciel, intégration, optimisation (chapitre 8).
- b) Programmation en **R** (chapitre 6).

Mises en relief

Nous avons souhaité soigner le mode de présentation de l'ouvrage (la forme) pour que l'information (le contenu) soit digeste. Par conséquent, des encadrés qui permettent la mise en relief de certaines informations importantes afin de faciliter la compréhension des notions abordées sont disposés à plusieurs endroits stratégiques du livre. Ces encadrés se distinguent par des icônes apparaissant dans la marge.

Astuce

Information supplémentaire relative au sujet traité.



Attention

Souligne un point important à ne pas négliger.



Remarque

Propose conseils et trucs pratiques.



Renvoi

Fait référence à un autre chapitre ou à un site internet.



Expert

Éléments avancés dont la lecture peut être omise en premier lieu.



Linux

Information réservée aux utilisateurs Linux.



Mac

Information réservée aux utilisateurs Macintosh.



Solutions des exercices et des travaux pratiques

Les corrigés des exercices et des séances de travaux pratiques sont fournis sur le site internet associé au livre (<http://www.biostatisticien.eu/springerR>).

Par ailleurs, quelques projets plus ambitieux que les travaux pratiques seront rendus accessibles sur ce site.

Conventions de police

- La lettre **R** désigne le logiciel **R**.
- Nous utiliserons l'écriture *italique* pour désigner des termes empruntés à la langue anglaise comme *data.frame* ou *package* ou bien des termes latins comme *versus* ou *a priori*.
- Nous utiliserons une police de caractères à **chasse fixe** (environnement **Verbatim**) pour noter des instructions **R**.
- Nous utiliserons une police de caractères en **PETITES CAPITALES** pour désigner un jeu de données et une police avec des **caractères sans empattement** pour désigner le nom du fichier physique contenant ce jeu de données. Cette dernière police de caractères sera utilisée pour indiquer n'importe quel fichier ou dossier mentionné dans cet ouvrage.

Sommaire

Avant-propos	ix
Liste des figures	xxix
Liste des tableaux	xxxiii
Notations mathématiques	xxxv
A Présentation du logiciel R	1
A.1 Présentation du logiciel	1
A.1.1 Origines	1
A.1.2 Pourquoi utiliser R ?	1
A.2 R et les statistiques	3
A.3 R et les graphiques	4
A.4 L'interface graphique de R (GUI)	5
A.5 Mes premiers pas en R	6
A.5.1 Utilisation de RCommander	6
A.5.1.1 Lancement de RCommander	6
A.5.1.2 Manipulation de données avec RCommander	8
A.5.1.3 Quelques manipulations statistiques avec RCom-	
mander	13
A.5.1.4 Rajouter des fonctionnalités à l'interface de	
RCommander	19
A.5.2 Utiliser R via la console	20
A.5.2.1 La force de R illustrée sur un exemple . . .	21
A.5.2.2 Un survol de la syntaxe de R via des com-	
mandes à taper	25
B Quelques jeux de données et problématiques	31
B.1 Indice de masse corporelle (IMC) chez des enfants	31
B.2 Poids de naissance	32
B.3 Épaisseur de l'intima-média	33
B.4 Alimentation chez des personnes âgées	34

B.5	Étude cas témoins sur l'infarctus du myocarde	35
B.6	Tableau résumant l'utilisation des jeux de données	36
I	Les bases du logiciel R	37
1	Les concepts de base, l'organisation des données	39
1.1	Votre première session	39
1.1.1	R est une calculatrice	40
1.1.2	Affichage des résultats et redirection dans des variables	41
1.1.3	Stratégie de travail	43
1.1.4	Utilisation de fonctions	47
1.2	Les données dans R	50
1.2.1	Nature (ou type, ou mode) des données	50
1.2.1.1	Type numérique (<code>numeric</code>)	50
1.2.1.2	† Type complexe (<code>complex</code>)	51
1.2.1.3	Type booléen ou logique (<code>logical</code>)	52
1.2.1.4	Données manquantes (<code>NA</code>)	52
1.2.1.5	Type chaînes de caractères (<code>character</code>)	53
1.2.1.6	† Données brutes (<code>raw</code>)	54
Récapitulatif	54	
1.2.2	Structures de données	55
1.2.2.1	Les vecteurs (<code>vector</code>)	55
1.2.2.2	Les matrices (<code>matrix</code>), les tableaux (<code>arrays</code>)	56
1.2.2.3	Les listes (<code>list</code>)	58
1.2.2.4	Le tableau individus × variables (<code>data.frame</code>)	59
1.2.2.5	Les facteurs (<code>factor</code>) et les variables ordinales (<code>ordered</code>)	60
1.2.2.6	Les dates	62
1.2.2.7	Les séries temporelles	62
Récapitulatif	63	
Termes à retenir	64	
Exercices	64	
Fiche de TP	65	
2	Importation-exportation et production de données	67
2.1	Importer des données	67
2.1.1	Importer des données depuis un fichier texte ASCII	67
2.1.1.1	Lecture de données avec <code>read.table()</code>	68
2.1.1.2	Lecture de données avec <code>read.ftable()</code>	71
2.1.1.3	Lecture de données avec la fonction <code>scan()</code>	72
2.1.2	Importer des données depuis Excel ou le tableur d'Open-Office	73
2.1.2.1	Utiliser le copier-coller	73

2.1.2.2	Passer par un fichier ASCII intermédiaire .	74
2.1.2.3	Utiliser des <i>packages</i> spécialisés	74
2.1.3	Importer des données depuis SPSS, Minitab, SAS ou Matlab	75
2.1.4	Les gros fichiers de données	75
2.2	Exporter des données	77
2.2.1	Exporter des données vers un fichier texte ASCII . .	77
2.2.2	Exporter des données vers Excel ou OpenOffice Calc	77
2.3	Création de données	77
2.3.1	Entrer des données jouets	77
2.3.2	Générer des données pseudo-aléatoires	79
2.3.3	Entrer des données issues d'un support papier . . .	79
2.4	† Lecture/écriture dans les bases de données	81
2.4.1	Créer une base de données et une table	81
2.4.2	Créer une source de données compatible avec MySQL	82
2.4.3	Écrire dans une table	83
2.4.4	Lire dans une table	84
	Termes à retenir	85
	Exercices	85
	Fiche de TP	86
3	Manipulation de données, fonctions	91
3.1	Opérations sur les vecteurs, matrices et listes	91
3.1.1	Arithmétique vectorielle	91
3.1.2	Le recyclage	92
3.1.3	Fonctions basiques	93
3.1.4	Opérations sur les matrices ou les <i>data.frames</i> . . .	94
3.1.4.1	Informations sur l'architecture	94
3.1.4.2	Fusion de tables	95
3.1.4.3	La fonction <code>apply()</code>	99
3.1.4.4	La fonction <code>sweep()</code>	100
3.1.4.5	La fonction <code>stack()</code>	100
3.1.4.6	La fonction <code>aggregate()</code>	101
3.1.4.7	La fonction <code>transform()</code>	102
3.1.5	Opérations sur les listes	102
3.2	Opérations logiques et relationnelles	103
3.3	Opérations ensemblistes	105
3.4	Extraction et insertion d'éléments	106
3.4.1	Extraction/Insertion dans les vecteurs	106
3.4.2	Extraction/Insertion dans les matrices	108
3.4.3	Extraction/Insertion dans les <i>arrays</i>	112
3.4.4	Extraction/Insertion dans les listes	113
3.5	Manipulation de chaînes de caractères	116

3.6	Manipulation de dates et d'unités de temps	119
3.6.1	Affichage de la date courante	119
3.6.2	Extraction de dates	119
3.6.3	Opérations sur des dates	121
3.7	Structures de contrôle	123
3.7.1	Instructions de condition	124
3.7.2	Instructions de boucles	127
3.8	Création de fonctions	129
3.9	† Représentation des nombres à virgule fixe, flottante	136
3.9.1	Représentation d'un nombre à l'aide d'une base	137
3.9.2	Représentation à virgule flottante	138
3.9.2.1	Définitions	138
3.9.2.2	Limite de cette représentation due à la man- tisse	139
3.9.2.3	Éviter certaines chausse-trappes numériques	140
3.9.2.4	Limite de cette représentation due à l'expo- sant	142
	Termes à retenir	145
	Exercices	145
	Fiche de TP	147
4	R et sa documentation	153
4.1	Aide intégrée au logiciel R	153
4.1.1	La commande <code>help()</code>	153
4.1.2	Quelques commandes complémentaires	155
4.2	† Aide accessible sur l'Internet	157
4.2.1	Moteurs de recherche	158
4.2.2	Forums de discussion	158
4.2.3	Listes de diffusion (<i>mailing lists</i>)	158
4.2.4	Discussion relayée par l'Internet (IRC)	159
4.2.5	<i>Wiki</i>	159
4.3	† Littérature sur R	159
4.3.1	Sur le web	159
4.3.2	En format papier	160
	Termes à retenir	161
	Exercices	161
	Fiche de TP	161
5	Techniques pour tracer des courbes et des graphiques	163
5.1	Les fenêtres graphiques	163
5.1.1	Fenêtre graphique de base, manipulation, sauvegarde	163
5.1.2	Découpage de la fenêtre graphique : <code>layout()</code>	165
5.2	Les fonctions de tracé de bas niveau	168

5.2.1	Les fonctions <code>plot()</code> et <code>points()</code>	168
5.2.2	Les fonctions <code>segments()</code> , <code>lines()</code> et <code>abline()</code> . .	170
5.2.3	La fonction <code>arrows()</code>	172
5.2.4	La fonction <code>polygon()</code>	173
5.2.5	La fonction <code>curve()</code>	173
5.2.6	La fonction <code>box()</code>	174
5.3	La gestion des couleurs	175
5.3.1	La fonction <code>colors()</code>	175
5.3.2	Le codage hexadécimal des couleurs	176
5.3.3	La fonction <code>image()</code>	179
5.4	L'ajout de texte	181
5.4.1	La fonction <code>text()</code>	181
5.4.2	La fonction <code>mtext()</code>	182
5.5	Titres, axes et légendes	183
5.5.1	La fonction <code>title()</code>	183
5.5.2	La fonction <code>axis()</code>	185
5.5.3	La fonction <code>legend()</code>	186
5.6	L'interaction avec le graphique	187
5.6.1	La fonction <code>locator()</code>	187
5.6.2	La fonction <code>identify()</code>	188
5.7	† La gestion fine des paramètres graphiques : <code>par()</code>	188
5.8	† Graphiques avancés : <code>rgl</code> , <code>lattice</code> et <code>ggplot2</code>	200
	Termes à retenir	202
	Exercices	202
	Fiche de TP	204
6	Programmation en R	209
6.1	Préambule	209
6.2	Développer des fonctions	210
6.2.1	Mise en route rapide : déclaration, création et appel de fonctions	210
6.2.2	Concepts de base sur les fonctions	211
6.2.2.1	Corps de fonction	211
6.2.2.2	Liste de paramètres formels et effectifs . .	211
6.2.2.3	Objet retourné par une fonction	215
6.2.2.4	Portée des variables dans le corps de la fonc- tion	217
6.2.3	Application à la problématique	219
6.2.4	Opérateurs	220
6.2.5	Le R vu comme un langage fonctionnel	222
6.3	† Programmation orientée objets	223
6.3.1	Comment fonctionne le mécanisme orienté objet du R	223
6.3.1.1	Classe d'un objet et déclaration d'un objet	223

6.3.1.2	Déclaration et utilisation d'une méthode d'un objet	224
6.3.2	Retour à la problématique	228
6.3.3	Information sur les méthodes	230
6.3.4	Héritage de classe	232
6.4	† Aller plus loin en programmation R	236
6.4.1	Attributs R	236
6.4.1.1	Attribut <code>class</code>	237
6.4.1.2	Attribut <code>dim</code>	238
6.4.1.3	Attributs <code>names</code> et <code>dimnames</code>	241
6.4.2	Autres objets R	244
6.4.2.1	Expression R	244
6.4.2.2	Formule R	247
6.4.2.3	Environnement R	249
6.5	† Interfacer R et C/C++ ou Fortran	251
6.5.1	Création et exécution d'une fonction C/C++ ou Fortran	253
6.5.2	Appel du code C/C++ (ou Fortran) depuis R	260
6.5.3	Appel de bibliothèques C/C++ ou Fortran externes	265
6.5.3.1	L'API R	266
6.5.3.2	La bibliothèque <code>newmat</code>	269
6.5.3.3	Les bibliothèques BLAS et LAPACK	271
6.5.3.4	Mélanger des bibliothèques C/C++ et Fortran	274
6.5.4	Appel d'un code R depuis un programme C/C++ appelé par R	276
6.5.5	Appel d'un code R depuis un programme Fortran	278
6.5.6	Quelques fonctions utiles	278
6.6	† Débogage de fonctions	279
6.6.1	Débogage de fonctions en R pur	279
6.6.2	Erreur dans le code R	281
6.6.3	Erreur dans le code C/C++ ou Fortran	282
6.6.4	Débogage avec GDB	283
6.6.4.1	Débogage avec Emacs	286
6.6.4.2	Débogage avec DDD	289
6.6.4.3	Débogage avec <code>Insight</code>	290
6.6.4.4	Détection de fuites de mémoire	294
6.7	Calcul parallèle et calculs sur cartes graphiques	297
6.7.1	Calcul parallèle	297
6.7.2	Calcul sur cartes graphiques	299
	Termes à retenir	301
	Exercices	301
	Fiche de TP	303

7	Maintenance des sessions	309
7.1	Les commandes R , les objets et leur stockage	309
7.2	Environnement de travail : les fichiers d'extension <code>.RData</code> .	311
7.3	Historique des commandes : les fichiers d'extension <code>.Rhistory</code>	314
7.4	Sauvegarder des graphiques	315
7.5	La gestion des <i>packages</i>	316
7.6	La gestion des chemins d'accès aux objets R	317
7.7	† Autres commandes utiles	319
7.8	† La gestion de la mémoire	320
7.8.1	Organisation de la mémoire vive	321
7.8.2	Accéder à la mémoire	321
7.8.2.1	Problèmes causés par la gestion mémoire des entiers	322
7.8.2.2	Allocation consécutive de la mémoire . . .	324
7.8.3	Taille des objets dans R	326
7.8.4	Quantité totale de mémoire utilisée par R	327
7.8.5	Quelques recommandations	329
7.9	† Utiliser R en mode <code>BATCH</code>	331
7.10	† Création d'un <i>package</i> R simplifié	332
	Termes à retenir	335
	Exercices	335
	Fiche de TP	336
II	Mathématiques et statistiques élémentaires	339
8	Mathématiques de base : calcul matriciel, intégration, optimisation	341
8.1	Les fonctions mathématiques de base	342
8.2	Calcul matriciel	343
8.2.1	Opérations de base	344
8.2.2	Produit extérieur	346
8.2.3	Produit de Kronecker	347
8.2.4	Matrices triangulaires	347
8.2.5	Opérateurs vec et demi-vec	348
8.2.6	Déterminant, trace, nombre de conditionnement . .	348
8.2.7	Données centrées, données réduites	349
8.2.8	Calcul des valeurs propres et vecteurs propres . . .	350
8.2.9	Racine carrée d'une matrice hermitienne définie positive	350
8.2.10	Décomposition en valeurs singulières	351
8.2.11	Décomposition de Cholesky	352
8.2.12	Décomposition QR	353

8.3	Intégration numérique	353
8.4	Dérivation	354
8.4.1	Dérivation symbolique	354
8.4.2	Dérivation numérique	355
8.5	Optimisation	356
8.5.1	Fonctions d'optimisation	356
8.5.2	Racines d'une fonction	360
	Termes à retenir	361
	Exercices	361
	Fiche de TP	362
9	Statistique descriptive	367
9.1	Introduction	367
9.2	Structuration des variables suivant leur type	368
9.2.1	Structurer les variables qualitatives	369
9.2.2	Structurer les variables ordinales	371
9.2.3	Structurer les variables quantitatives discrètes	371
9.2.4	Structurer les variables quantitatives continues	371
9.3	Tableaux de données	372
9.3.1	Tableaux des données individuelles	372
9.3.2	Tableaux des effectifs ou des fréquences d'une variable	372
9.3.3	Tableaux de données regroupées en classes	373
9.3.4	Tableaux croisant deux variables	373
9.3.4.1	Tableaux de contingence	373
9.3.4.2	Distribution conjointe	374
9.3.4.3	Distributions marginales	375
9.3.4.4	Distributions conditionnelles	375
9.4	Résumés numériques	376
9.4.1	Résumés de position d'une distribution	377
9.4.1.1	Le (ou les) mode(s)	377
9.4.1.2	La médiane	377
9.4.1.3	La moyenne	379
9.4.1.4	Les fractiles	379
9.4.2	Résumés de dispersion d'une distribution	380
9.4.3	Résumés de forme d'une distribution	381
9.5	Mesures d'association	381
9.5.1	Mesures de liaison entre deux variables qualitatives	381
9.5.1.1	La statistique du χ^2 de Pearson	381
9.5.1.2	Φ^2 , V de Cramér et coefficient de contingence de Pearson	382
9.5.2	Mesures de liaison entre des variables ordinales (ou des rangs)	383
9.5.2.1	Le τ et le τ_b de Kendall	383

9.5.2.2	Coefficient ρ de corrélation des rangs de Spearman	384
9.5.3	Mesures de liaison entre deux variables quantitatives	385
9.5.3.1	Covariance et coefficient de corrélation de Pearson	385
9.5.4	Mesures de liaison entre une variable quantitative et une variable qualitative	385
9.5.4.1	Le rapport de corrélation $\eta^2_{Y X}$	385
9.6	Représentations graphiques	386
9.6.1	Graphiques pour les variables qualitatives	387
9.6.1.1	Diagramme en croix	387
9.6.1.2	Diagramme en tuyaux d'orgue	388
9.6.1.3	Diagramme de Pareto	389
9.6.1.4	Diagramme empilé	390
9.6.1.5	Diagramme circulaire	391
9.6.2	Graphiques pour les variables ordinales	392
9.6.2.1	Diagramme en tuyaux d'orgue avec courbe des fréquences cumulées	392
9.6.3	Graphiques pour les variables quantitatives discrètes	392
9.6.3.1	Diagramme en croix	392
9.6.3.2	Diagramme en bâtons	393
9.6.3.3	Graphe de la fonction de répartition empirique	393
9.6.3.4	Diagramme en tiges et feuilles	394
9.6.3.5	Boîte à moustaches (<i>boxplot</i>)	394
9.6.4	Graphiques pour les variables quantitatives continues	396
9.6.4.1	Graphe de la fonction de répartition empirique	396
9.6.4.2	Diagramme en tiges et feuilles	397
9.6.4.3	Boîte à moustaches	398
9.6.4.4	Histogramme en densité à amplitudes de classes égales ou inégales	398
9.6.4.5	Polygone des fréquences	400
9.6.4.6	Polygone des fréquences cumulées	400
9.6.5	Représentations graphiques dans un cadre bivarié .	401
9.6.5.1	Croisement de deux variables qualitatives .	401
9.6.5.2	Croisement de deux variables quantitatives	404
9.6.5.3	Croisement d'une variable qualitative et d'une variable quantitative	405
	Termes à retenir	406
	Exercices	406
	Fiche de TP	407

10	Variables aléatoires, lois et simulations : une meilleure compréhension grâce aux spécificités de R	411
10.1	Notions sur la génération de nombres au hasard	411
10.2	La notion de variable aléatoire	413
10.2.1	Réalisations d'une variable aléatoire et loi de fonction- nement	413
10.2.2	Variables aléatoires <i>i.i.d.</i>	415
10.2.3	Caractériser la loi d'une variable aléatoire	416
10.2.3.1	Densité, fonction de répartition, fonction quan- tile	417
10.2.4	Paramètres de la loi d'une variable aléatoire	420
10.3	Loi des grands nombres et théorème de la limite centrale	423
10.3.1	Loi des grands nombres	423
10.3.2	Théorème de la limite centrale	424
10.4	La statistique inférentielle	425
10.4.1	Estimation (ponctuelle) de paramètres	425
10.4.2	La fonction de répartition empirique	427
10.4.3	Estimation par la méthode du maximum de vraisem- blance	428
10.4.4	Fluctuation d'échantillonnage et qualités d'un estima- teur	430
10.5	Quelques techniques de simulation (d'une loi)	432
10.5.1	Simuler à partir d'une autre loi	433
10.5.2	Méthode de la transformation inverse	433
10.5.3	Méthode du rejet	434
10.5.4	Simulation de variables aléatoires discrètes	435
10.6	La méthode du <i>bootstrap</i>	435
10.7	Lois usuelles et moins usuelles	436
10.7.1	Lois usuelles	436
10.7.2	† Lois moins usuelles	439
10.8	Modélisation d'un phénomène	440
	Termes à retenir	444
	Exercices	444
	Fiche de TP	444
11	Intervalles de confiance et tests d'hypothèses	449
11.1	Notations	449
11.2	Intervalles de confiance	450
11.2.1	Intervalles de confiance pour une moyenne	451
11.2.2	Intervalles de confiance pour une proportion	452
11.2.3	Intervalles de confiance pour une variance	453
11.2.4	Intervalles de confiance pour une médiane	455

11.2.5	Intervalle de confiance pour un coefficient de corrélation	456
11.2.6	Tableau récapitulatif des intervalles de confiance	456
11.3	Tests d'hypothèses usuels	457
11.3.1	Tests paramétriques	459
11.3.1.1	Tests de moyenne	459
11.3.1.2	Tests de variance	462
11.3.1.3	Tests de proportion	464
11.3.1.4	Tests de coefficient de corrélation	467
11.3.2	Tests d'indépendance	468
11.3.2.1	Test du χ^2 d'indépendance	468
11.3.2.2	Test du χ^2 de Yates	470
11.3.2.3	Test de Fisher exact	471
11.3.3	Tests non paramétriques	472
11.3.3.1	Tests d'adéquation	472
11.3.3.2	Tests de position	476
11.3.4	Tableau récapitulatif des tests usuels	481
11.4	Autres tests d'hypothèses	481
	Termes à retenir	483
	Exercices	483
	Fiche de TP	484

12 Régression linéaire simple et multiple 489

12.1	Introduction	489
12.2	La régression linéaire simple	491
12.2.1	Objectif et modèle	491
12.2.2	Ajustement sur des données	491
12.2.3	Intervalle de confiance et de prédiction pour une nouvelle valeur	496
12.2.4	Analyse des résidus	499
12.2.5	Tests de Student pour des moyennes et modèle linéaire	502
12.2.6	Récapitulatif	503
12.3	La régression linéaire multiple	504
12.3.1	Objectif et modèle	504
12.3.2	Ajustement sur des données	504
12.3.3	Intervalle de confiance et de prédiction pour une nouvelle valeur	509
12.3.4	Test d'une sous-hypothèse linéaire : test de Fisher partiel	509
12.3.5	Cas des variables qualitatives à plus de deux modalités	510
12.3.6	Interaction entre les variables	514
12.3.7	Problème de la colinéarité	518
12.3.8	Sélection de variables	519
12.3.9	Analyse des résidus	528

12.3.10	Cas de la régression polynomiale	535
12.3.11	Récapitulatif	535
	Termes à retenir	536
	Exercices	536
	Fiche de TP	537
13	Analyse de variance élémentaire	541
13.1	Analyse de la variance à un facteur	541
13.1.1	Les objectifs, les données et le modèle	541
13.1.2	Exemple et inspection graphique	542
13.1.3	Table d'ANOVA et estimations des paramètres . . .	544
13.1.4	Validation des hypothèses	547
13.1.5	Comparaisons multiples et contrastes	548
13.1.6	Récapitulatif	551
13.2	Analyse de la variance à deux facteurs	552
13.2.1	Objectifs, données et modèle	552
13.2.2	Exemple et inspection graphique	553
13.2.3	Table d'ANOVA, tests et estimation des paramètres	555
13.2.4	Validation des hypothèses	558
13.2.5	Contrastes	559
13.2.6	Récapitulatif	560
13.3	Analyses de variance à mesures répétées	561
13.3.1	Modèle à un facteur à mesures répétées	562
13.3.2	Modèle à deux facteurs à mesures répétées sur les deux facteurs	563
13.3.3	Modèle à deux facteurs à mesures répétées sur un seul facteur	565
	Termes à retenir	567
	Exercices	567
	Fiche de TP	567
	Annexes : Installation du logiciel R et des <i>packages</i> R	573
C.1	Installation de R sous Microsoft Windows	573
C.2	Installation de <i>packages</i> supplémentaires	574
C.2.1	Installation à partir d'un fichier situé sur le disque .	574
C.2.2	Installation directement depuis l'Internet	575
C.2.3	Installation depuis la ligne de commande	576
C.2.4	Installation de <i>packages</i> sous Linux	577
C.3	Chargement des <i>packages</i> installés	578
	Références	581
	Index général	585

Index des commandes et des symboles R	595
Index des auteurs	609
Liste des <i>packages</i> R mentionnés dans le livre	611
Solutions des exercices	613
Solutions des TPs	625

Liste des figures

A.1	Quelques possibilités graphiques offertes par R	5
A.2	L'interface graphique de RCommander	8
A.3	Entrer des données via l'interface graphique de RCommander	9
A.4	Statistiques élémentaires avec RCommander	11
A.5	Manipulation d'un jeu de données avec RCommander	12
A.6	Test de moyennes avec RCommander	15
A.7	Test d'indépendance avec RCommander	17
A.8	Plan des moindres carrés.	19
1.1	Vue de la fenêtre de script et de la console de commandes.	44
1.2	Caractéristiques d'un nombre complexe.	51
1.3	Illustration d'une <i>array</i>	57
5.1	Effet du paramètre <code>mfrow</code> de la fonction <code>par()</code>	166
5.2	Visualisation du potentiel de la fonction <code>layout()</code>	167
5.3	La fonction <code>layout()</code> et ses paramètres <code>widths</code> et <code>heights</code>	168
5.4	La fonction <code>plot()</code>	169
5.5	La fonction <code>points()</code>	170
5.6	Les fonctions <code>segments()</code> et <code>lines()</code>	171
5.7	La fonction <code>abline()</code>	171
5.8	La fonction <code>arrows()</code>	172
5.9	La fonction <code>curve()</code>	173
5.10	La fonction <code>box()</code>	174
5.11	Le paramètre <code>col</code> de la fonction <code>plot()</code>	175
5.12	Le paramètre <code>alpha</code> de la fonction <code>rgb()</code>	177
5.13	Un exemple utilisant la fonction <code>rainbow()</code>	178
5.14	La fonction <code>display.brewer.all()</code>	179
5.15	La fonction <code>image()</code>	180
5.16	La fonction <code>image()</code> , affichage cohérent avec les données.	181
5.17	La fonction <code>text()</code>	182
5.18	La fonction <code>mtext()</code>	183
5.19	La fonction <code>title()</code>	184
5.20	Titre sur plusieurs lignes dans un graphique.	184

5.21	La fonction <code>axis()</code>	185
5.22	La fonction <code>legend()</code> avec des carrés.	186
5.23	La fonction <code>legend()</code> avec des segments.	187
5.24	Figure illustrant la gestion fine des paramètres graphiques.	191
5.25	Gestion des couleurs sur un graphique.	192
5.26	Mise en situation des paramètres <code>adj</code> et <code>srt</code>	194
5.27	Utiliser diverses polices sur un graphique.	195
5.28	Gestion des étiquettes sur un graphique.	197
5.29	Les paramètres <code>lend</code> et <code>ljoin</code>	199
5.30	Le paramètre <code>pch</code>	199
5.31	Les paramètres <code>lty</code> et <code>lwd</code>	200
6.1	Résultat de l'appel de la fonction <code>affiche.reg1()</code>	220
6.2	Emacs et GDB.	288
6.3	DDD et GDB.	290
7.1	Stockage de valeurs dans la mémoire.	321
7.2	Stockage par R d'un <i>integer</i> (signé) dans la mémoire.	322
8.1	Fonction sinc modifiée.	357
9.1	Algorithme de détermination du type d'une variable.	369
9.2	Diagramme en croix pour une variable qualitative.	387
9.3	Diagramme en points pour une variable qualitative.	388
9.4	Diagramme en tuyaux d'orgue pour une variable qualitative.	388
9.5	Diagramme de Pareto pour une variable qualitative.	389
9.6	Diagramme empilé pour une variable qualitative.	390
9.7	Tuyaux d'orgue pour une variable ordinale.	392
9.8	Diagramme en bâtons pour une variable quantitative discrète.	393
9.9	Fonction de répartition empirique pour une variable discrète.	394
9.10	Boîte à moustaches et explications associées.	396
9.11	Fonction de répartition empirique pour une variable continue.	397
9.12	Histogramme à amplitudes de classes égales ou inégales.	399
9.13	Polygone des fréquences.	400
9.14	Polygone des fréquences cumulées.	401
9.15	Tuyaux d'orgue pour deux variables qualitatives.	402
9.16	Diagramme mosaïque pour deux variables qualitatives.	402
9.17	Graphique de Cohen-Friendly pour variables qualitatives.	403
9.18	Graphique <code>table.cont</code> croisant deux variables qualitatives.	403
9.19	Graphique croisant deux variables quantitatives.	404
9.20	<i>Boxplots</i> d'une variable quantitative, niveaux d'un facteur.	405
9.21	<code>stripchart</code> : croiser variable quantitative et qualitative.	405
10.1	Courbe approchant la densité de X.	419
10.2	Convergence en loi en action, données simulées.	425

12.1	Nuage de points du poids de l'enfant <i>vs</i> celui de la mère. . .	492
12.2	Droite de régression des moindres carrés.	493
12.3	Intervalle de confiance et intervalle de prévision.	498
12.4	Inspection graphique de la normalité des résidus.	500
12.5	Graphe des résidus en fonction des valeurs prédites.	501
12.6	Diagramme de dispersion de toutes les paires de variables.	506
12.7	Effet de l'âge sur BWT dans un modèle sans interaction. . .	516
12.8	Effet de l'âge sur BWT dans un modèle avec interaction. . .	517
12.9	Sélection de variables par le critère BIC.	521
12.10	Inspection de l'hypothèse d'homoscédasticité et de normalité.	528
12.11	Résidus en fonction des variables explicatives.	529
12.12	Points atypiques : résidus studentisés <i>versus</i> valeurs ajustées.	531
12.13	Visualisation d'observations influentes : distance de Cook. .	533
13.1	Boîtes à moustaches des délais de cicatrisation par traitement.	544
13.2	Analyser les résidus dans une ANOVA à un facteur.	547
13.3	Interaction dans une ANOVA à deux facteurs.	554
13.4	Analyser les résidus dans une ANOVA à deux facteurs. . .	559

Liste des tableaux

1.1	Les différents types de données en R	54
1.2	Les différentes structures de données en R	63
2.1	Fonctions d'importation de données.	68
2.2	Paramètres principaux de <code>read.table()</code>	68
2.3	<i>Packages</i> et fonctions R d'importation de données.	75
3.1	Opérateurs et fonctions agissant sur ou créant des logiques.	104
3.2	Opérations ensemblistes.	105
3.3	Codes pour la fonction <code>strptime()</code>	120
3.4	Correspondance entre IMC et types de corpulence.	133
5.1	Paramètres de gestion de la fenêtre graphique.	190
5.2	Paramètres de gestion de la couleur.	192
5.3	Paramètres de gestion du texte affiché sur le graphique.	193
5.4	Paramètres pour la gestion des axes.	196
5.5	Paramètres pour la gestion des lignes et symboles.	198
6.1	Conventions sur les types des arguments.	261
8.1	Tableau des fonctions mathématiques de base.	342
10.1	Lois discrètes usuelles.	437
10.2	Lois continues usuelles.	438
10.3	Lois moins usuelles I.	439
10.4	Lois moins usuelles II.	440
11.1	Notations sur les estimations de paramètres classiques.	450
11.2	Notation des différents quantiles d'ordre p	450
11.3	Résumé sur les intervalles de confiance.	456
11.4	Les tests usuels.	481
12.1	Principales fonctions R en régression linéaire simple.	503
12.2	Principales fonctions R en régression linéaire multiple.	535

13.1	Principales fonctions à utiliser en ANOVA à un facteur. . .	551
13.2	Principales fonctions pour une ANOVA à deux facteurs. . .	560

Notations mathématiques

$:=$	Symbole indiquant des notations différentes pour un même objet
\cup	Fusion de tables
$a \in A$	a appartient à l'ensemble A
$A \subset B$	A inclus dans B
$A \supset B$	A contient B
$A \cap B$	Intersection des ensembles A et B
$A \cup B$	Réunion des ensembles A et B
$A \setminus B$	Complémentaire de l'ensemble B dans l'ensemble A
$(A \cup B) \setminus (A \cap B)$	Différence symétrique des ensembles A et B
f_i	Fréquence d'une modalité
$ x $	Valeur absolue du nombre x
$x!$	Factorielle du nombre x
$\binom{n}{p}$	Nombre de combinaisons de p éléments pris parmi n , coefficients du binôme
$\Gamma(\cdot)$	Fonction gamma
γ	Constante d'Euler
$\psi(\cdot)$	Fonction digamma
π	Nombre π
λ	Nombre scalaire
$\mathcal{A}, \mathcal{B}, \mathcal{C}$, etc.	Matrices
I	Matrice identité
$n \times p$	Pour indiquer la taille d'une matrice
\mathcal{A}^\top	Transposée de la matrice \mathcal{A}
\mathcal{B}^{-1}	Inverse de la matrice \mathcal{B}
$\overline{\mathcal{C}}$	Conjuguée de la matrice complexe \mathcal{C}
$\mathbf{x} = (x_1, \dots, x_n)^\top$	Vecteur d'éléments en colonne
\mathbf{x}^\top	Transposée du vecteur \mathbf{x}
$\mathcal{A} \otimes \mathcal{B}$	Produit de Kronecker de la matrice \mathcal{A} par la matrice \mathcal{B}
$\text{vec}(\mathcal{A})$	Vecteur de l'empilement des colonnes de la matrice \mathcal{A}
$\text{vech}(\mathcal{A})$	Vecteur de l'empilement des colonnes de la matrice \mathcal{A} , mais en excluant les éléments au-dessus de la diagonale
\mathcal{M}^*	Matrice adjointe (transposée conjuguée) de la matrice \mathcal{M}
$*$	Produit usuel

$\mathbf{M}^{1/2}$	Racine carrée de la matrice \mathbf{M}
$\mathbf{1}_{[A]}(x)$	Vaut 1 si $x \in A$ et 0 sinon
$[a, b]$	Intervalle des valeurs comprises entre a et b
$\det(\mathcal{A})$	Déterminant de la matrice \mathcal{A}
$\Phi(\cdot)$	Fonction de répartition d'une variable aléatoire de loi normale standard $\mathcal{N}(0, 1)$
$\dot{\mathcal{X}}$	Matrice obtenue en centrant les colonnes de la matrice \mathcal{X}
$\mathbf{1}_n$	Vecteur $(1, \dots, 1)^\top$ de longueur n
X, Y	Variables non aléatoires (statistique descriptive)
N	Taille de la population
n	Taille échantillonnale
$m_e := q_{1/2}$	Médiane
$PFC_X(\cdot)$	Valeur du polygone des fréquences cumulées de X
μ_X, μ	Espérance de la variable aléatoire X ou moyenne de la population en statistique descriptive
q_p ou x_p	Fractile (quantile) d'ordre p d'une variable
$q_{1/4}, q_{3/4}$	Premier et troisième quartile (aussi notés q_1 et q_3)
$\sigma_{Pop}^2(\mathbf{x})$	Variance de la population (statistique descriptive)
$\sigma_{Pop}(\mathbf{x})$	Écart type de la population (statistique descriptive)
c_v	Coefficient de variation de la population (statistique descriptive)
γ_1	Coefficient d'asymétrie (<i>skewness</i>)
β_2	Coefficient d'aplatissement (<i>kurtosis</i>)
μ_3	Moment centré d'ordre 3
μ_4	Moment centré d'ordre 4
χ^2	Statistique du χ^2 de Pearson
Φ^2, V^2	Φ^2 et V^2 de Cramér
τ, τ_b	τ et τ_b de Kendall
ρ	Coefficient de corrélation théorique de Pearson
$\eta_{Y X}^2$	Rapport de corrélation
X, Y, ϵ	Variables aléatoires
x_i, y_i, ϵ_i	Réalisations des variables aléatoires X, Y, ϵ
$\mathbf{X}, \mathbf{Y}, \epsilon$	Vecteurs aléatoires
\mathbf{X}_n	Échantillon (aléatoire)
x_n	Échantillon (observé)
\mathbf{X}	Matrice aléatoire
\mathcal{L}	Loi (générique) d'une variable aléatoire
$\mathcal{N}(0, 1)$	Loi gaussienne standard
$\mathcal{N}(\mu, \sigma^2)$	Loi gaussienne (normale) d'espérance μ et de variance σ^2
$\mathcal{U}(a, b)$	Loi uniforme sur l'intervalle $[a, b]$
$\text{Bin}(n, p)$	Loi binomiale de paramètres n et p

$\mathcal{E}(\lambda)$	Loi exponentielle de paramètre λ
$\mathcal{P}(\lambda)$	Loi de Poisson de paramètre λ
$\mathcal{T}(n)$	Loi de Student à n degrés de liberté
$\chi^2(n)$ ou χ_n^2	Loi du χ^2 à n degrés de liberté
$\mathcal{F}(n, m)$	Loi de Fisher à n et m degrés de liberté
$f_X(\cdot)$	Fonction de densité de la variable aléatoire X
$F_X(\cdot)$	Fonction de répartition de la variable aléatoire X
$F_X^{-1}(\cdot)$	Fonction de répartition réciproque de la variable aléatoire X
σ^2	Variance d'une variable aléatoire
$\mathbb{E}(Y)$	Espérance théorique de la variable aléatoire Y
$\text{Var}(Y)$	Variance théorique de la variable aléatoire Y
\bar{X}_n	Moyenne empirique $\frac{1}{n} \sum_{i=1}^n X_i$ de l'échantillon $\mathbf{X}_n = (X_1, \dots, X_n)^\top$, estimateur de μ_X
\bar{x}_n	Réalisation de la moyenne empirique $\frac{1}{n} \sum_{i=1}^n X_i$ de l'échantillon $\mathbf{X}_n = (X_1, \dots, X_n)^\top$, estimation de μ_X
\xrightarrow{P}	Symbole de convergence en probabilité
$\hat{F}_n(\cdot) := \hat{F}_{\mathbf{X}_n}(\cdot)$	Fonction de répartition empirique de l'échantillon \mathbf{X}_n
θ	Paramètre inconnu (parfois on notera θ^\bullet la vraie valeur inconnue du paramètre)
$\hat{\theta}(X_1, \dots, X_n)$ ou $\hat{\theta}$	Estimateur du paramètre inconnu θ basé sur l'échantillon $\mathbf{X}_n = (X_1, \dots, X_n)^\top$
$\hat{\theta}(x_1, \dots, x_n)$ ou $\hat{\theta}$	Estimation du paramètre inconnu θ basé sur l'échantillon observé $\mathbf{x}_n = (x_1, \dots, x_n)^\top$
$\mathbb{B}(\hat{\theta}(X_1, \dots, X_n); \theta)$	Biais de l'estimateur $\hat{\theta}(X_1, \dots, X_n)$ pour estimer le paramètre inconnu θ
$P[A]$	Probabilité de l'ensemble A
$\mathcal{V}(\theta; X_1, \dots, X_n)$	Fonction de vraisemblance de l'échantillon \mathbf{X}_n évaluée en θ
$\mathbf{x}^* = (x_1^*, \dots, x_n^*)^\top$	Échantillon <i>bootstrap</i> généré à partir de l'échantillon observé $\mathbf{x}_n = (x_1, \dots, x_n)^\top$
$\hat{\sigma}$	Estimateur de σ
$\hat{\sigma}$	Estimation de σ
p	Proportion théorique
\hat{p}	Estimateur d'une proportion (ou d'une probabilité)
\hat{p}	Estimation d'une proportion (ou d'une probabilité)
\widehat{m}_e	Estimateur d'une médiane
\widehat{m}_e	Estimation d'une médiane
M	Nombre de boucles (d'échantillons générés) dans une simulation de Monte-Carlo
B	Nombre d'échantillons <i>bootstrap</i> générés
$B(\cdot, \cdot)$	Fonction bêta

$I'_x(\cdot, \cdot)$	Dérivée de la fonction bêta incomplète
$I(\cdot)$	Fonction de Bessel modifiée
$I_\alpha(\cdot)$	Fonctions de Bessel modifiées
u_p	Quantile d'ordre p d'une $\mathcal{N}(0, 1)$
t_p^n	Quantile d'ordre p d'une $\mathcal{T}(n)$
q_p^n	Quantile d'ordre p d'une $\chi^2(n)$
$f_p^{n,m}$	Quantile d'ordre p d'une $\mathcal{F}(n, m)$
$IC_{1-\alpha}(\theta)$	Intervalle de confiance (aléatoire) de niveau de confiance $1 - \alpha$ pour θ
$ic_{1-\alpha}(\theta)$	Intervalle de confiance (réalisé) de niveau de confiance $1 - \alpha$ pour θ
$1 - \alpha$	Niveau de confiance d'un intervalle de confiance
$(x_{(1)}, \dots, x_{(n)})$	Échantillon (observé) ordonné par valeurs croissantes
\mathcal{H}_1	Assertion d'intérêt dans les tests d'hypothèses
\mathcal{H}_0	Hypothèse dite nulle, contraire de \mathcal{H}_1
α	Niveau de signification ou risque de première espèce dans les tests d'hypothèses
R	Coefficient de corrélation empirique aléatoire de Pearson
r	Coefficient de corrélation empirique réalisé de Pearson
β_0, β_1	Coefficients inconnus d'un modèle de régression linéaire simple
$\hat{\beta}_0, \hat{\beta}_1$	Estimations des coefficients inconnus d'un modèle de régression linéaire simple
$\hat{\epsilon}_i$	Résidus observés d'un modèle de régression linéaire simple
\hat{y}_i	Valeurs ajustées observées d'un modèle de régression linéaire simple
R^2	Coefficient de détermination aléatoire en régression
r^2	Coefficient de détermination réalisé en régression
R_a^2	Coefficient de détermination ajusté aléatoire en régression
r_a^2	Coefficient de détermination ajusté réalisé en régression
\hat{Y}^p	Préviseur de la variable aléatoire Y pour une nouvelle valeur de la variable explicative X en régression
$IP_{1-\alpha}(Y_0, x_0)$	Intervalle de prévision de niveau $1 - \alpha$ pour la variable aléatoire Y_0 associée à une nouvelle valeur x_0 de la variable explicative
$\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top$	Vecteur des $p + 1$ coefficients inconnus d'un modèle de régression linéaire multiple à p variables explicatives

$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$	Estimateur du vecteur des paramètres inconnus β pour la matrice observée des variables explicatives \mathbf{X} et pour le vecteur des valeurs à expliquer dans un modèle de régression linéaire multiple
$\hat{\beta}$	Estimation de β
VIF	Facteur d'inflation de la variance, <i>variance inflation factor</i>
AIC	<i>an information criterion</i>
BIC	<i>bayesian information criterion</i>
h_{ii}	Levier de la i -ème observation en régression
t_i	Résidus standardisés
t_i^*	Résidus studentisés
$\hat{\sigma}_{(-i)}$	Estimation de σ sans utiliser la i -ème observation
C_i	Distances de Cook
$\hat{y}_j^{(-i)}$	Prédiction de y_j sans utiliser la i -ème observation
$\hat{\beta}_j^{(-i)}$	Estimation de β_j sans utiliser la i -ème observation
I, J	Nombre de niveaux d'un facteur en ANOVA
$\mu_{..}$	Effet moyen général en ANOVA
$\mu_{i.}$	Effet du niveau i d'un facteur en ANOVA
$\mu_{.j}$	Effet du niveau j d'un facteur en ANOVA



Collection
**Statistique
et probabilités
appliquées**

**Dirigée par
Yadolah Dodge**

COMITÉ ÉDITORIAL:

Aurore Delaigle

Université de Melbourne, Australie

Christian Genest

Université Laval, Québec

Marc Hallin

Université libre de Bruxelles, Belgique

Ludovic Lebart

Télécom-ParisTech, Paris

Christian Mazza

Université de Fribourg, Suisse

Stephan Morgenthaler

EPFL, Lausanne

Louis-Paul Rivest

Université Laval, Québec

Gilbert Saporta

CNAM, Paris

**Pierre Lafaye de Micheaux,
Rémy Drouilhet, Benoît Liquet**

Le logiciel R

Deuxième
édition

Maîtriser le langage

Effectuer des analyses (bio)statistiques

Cette collection met à la disposition du public intéressé par la statistique (étudiants, enseignants, chercheurs) des ouvrages qui concilient effort pédagogique et travail permanent de mise à jour.

Cette démarche implique de prendre en compte de façon sélective et critique les renouvellements des concepts, des champs d'application et des outils de traitement. Seules une compréhension profonde et une appropriation des connaissances permettront de s'adapter aux évolutions qui n'ont pas fini de bouleverser cette discipline.

Ce livre est consacré à un outil désormais incontournable pour l'analyse de données, l'élaboration de graphiques et le calcul statistique : le logiciel R.

Après avoir introduit les principaux concepts permettant une utilisation sereine de cet environnement informatique (organisation des données, importation et exportation, accès à la documentation, représentations graphiques, programmation, maintenance, etc.), les auteurs de cet ouvrage détaillent l'ensemble des manipulations permettant la manipulation avec R d'un très grand nombre de méthodes et de notions statistiques : simulation de variables aléatoires, intervalles de confiance, tests d'hypothèses, valeur-p, bootstrap, régression linéaire, ANOVA (y compris répétées), et d'autres encore.

Écrit avec un grand souci de pédagogie et clarté, et agrémenté de nombreux exercices et travaux pratiques, ce livre accompagnera idéalement tous les utilisateurs de R – et ceci sur les environnements Windows, Macintosh ou Linux – qu'ils soient débutants ou d'un niveau avancé : étudiants, enseignants ou chercheurs en statistique, mathématique, médecine, informatique, biologie, psychologie, sciences infirmières, etc. Il leur permettra de maîtriser en profondeur le fonctionnement de ce logiciel. L'ouvrage sera aussi utile aux utilisateurs plus confirmés qui retrouveront exposées ici l'ensemble des fonctions R les plus couramment utilisées.



editions.lavoisier.fr