Collection
Statistique et probabilités appliquées

Pierre Lafaye de Micheaux
Rémy Drouilhet
Benoît Liquet

Le logiciel R

Maîtriser le langage
Effectuer des analyses (bio)statistiques

> lm(y-x)

> eigen(X)$val

> demo

> boxplot(x)

> apply (X,FUN=mean,MARGIN

> t.test(x)

> vecn <- function (n)

Lavoisier
Hermes

Le logiciel R

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

Pierre Lafaye de Micheaux
Rémy Drouilhet
Benoît Liquet

Le logiciel R

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

GQ@L ⁸ 756+0+5240+26/6+0

Imprimé en France

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 :

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, 2e 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, 2e 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 Raymondaud 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, un 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/springeR`).

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

Liste des figures

Liste des tableaux

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
$A \backslash B$	Complémentaire de l'ensemble B dans l'ensemble A
$(A \cup B) \backslash (A \cap B)$	Différence symétrique des ensembles A et B
f_i	Fréquence d'une modalité
$\lvert x \rvert$	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
$\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{C}}$, etc.	Matrices
$\boldsymbol{\mathcal{I}}$	Matrice identité
$n \times p$	Pour indiquer la taille d'une matrice
$\boldsymbol{\mathcal{A}}^\mathsf{T}$	Transposée de la matrice $\boldsymbol{\mathcal{A}}$
$\boldsymbol{\mathcal{B}}^{-1}$	Inverse de la matrice $\boldsymbol{\mathcal{B}}$
$\overline{\boldsymbol{\mathcal{C}}}$	Conjuguée de la matrice complexe $\boldsymbol{\mathcal{C}}$
$\boldsymbol{x} = (x_1, \ldots, x_n)^\mathsf{T}$	Vecteur d'éléments en colonne
$\boldsymbol{x}^\mathsf{T}$	Transposée du vecteur $\boldsymbol{x}$
$\boldsymbol{\mathcal{A}} \otimes \boldsymbol{\mathcal{B}}$	Produit de Kronecker de la matrice $\boldsymbol{\mathcal{A}}$ par la matrice $\boldsymbol{\mathcal{B}}$
$vec(\boldsymbol{\mathcal{A}})$	Vecteur de l'empilement des colonnes de la matrice $\boldsymbol{\mathcal{A}}$
$vech(\boldsymbol{\mathcal{A}})$	Vecteur de l'empilement des colonnes de la matrice $\boldsymbol{\mathcal{A}}$, mais en excluant les éléments au-dessus de la diagonale
$\boldsymbol{\mathcal{M}}^*$	Matrice adjointe (transposée conjuguée) de la matrice $\boldsymbol{\mathcal{M}}$
$\star$	Produit usuel

$\boldsymbol{\mathcal{M}}^{1/2}$	Racine carrée de la matrice $\boldsymbol{\mathcal{M}}$
$\mathbf{1}_{[A]}(x)$	Vaut 1 si $x \in A$ et 0 sinon
$[a, b]$	Intervalle des valeurs comprises entre a et b
$\det(\boldsymbol{\mathcal{A}})$	Déterminant de la matrice $\boldsymbol{\mathcal{A}}$
$\Phi(\cdot)$	Fonction de répartition d'une variable aléatoire de loi normale standard $\mathcal{N}(0, 1)$
$\dot{\boldsymbol{\mathcal{X}}}$	Matrice obtenue en centrant les colonnes de la matrice $\boldsymbol{\mathcal{X}}$
$\mathbf{1}_n$	Vecteur $(1, \ldots, 1)^\mathsf{T}$ 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)
$\sigma^2_{Pop}(\boldsymbol{x})$	Variance de la population (statistique descriptive)
$\sigma_{Pop}(\boldsymbol{x})$	Écart type de la population (statistique descriptive)
c_v	Coefficient de variation de la population (statistique descriptive)
γ_1	Coefficient d'asymétrie (*skewness*)
β_2	Coefficient d'aplatissement (*kurtosis*)
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
$\eta^2_{Y\|X}$	Rapport de corrélation
X, Y, ϵ	Variables aléatoires
x_i, y_i, ϵ_i	Réalisations des variables aléatoires X, Y, ϵ
$\mathbf{X}$, $\mathbf{Y}$, $\boldsymbol{\epsilon}$	Vecteurs aléatoires
$\mathbf{X}_n$	Échantillon (aléatoire)
$\boldsymbol{x}_n$	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 $[a, b]$
$\mathcal{B}in(n, p)$	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^2_n	Loi du χ^2 à n degrés de liberté
$\mathcal{F}(n,m)$	Loi de Fisher à n et m degrés de liberté
$f_{\mathrm{X}}(\cdot)$	Fonction de densité de la variable aléatoire X
$F_{\mathrm{X}}(\cdot)$	Fonction de répartition de la variable aléatoire X
$F_{\mathrm{X}}^{-1}(\cdot)$	Fonction de répartition réciproque de la variable aléatoire X
σ^2	Variance d'une variable aléatoire
$\mathbb{E}(\mathrm{Y})$	Espérance théorique de la variable aléatoire Y
$\mathbb{V}\mathrm{ar}(\mathrm{Y})$	Variance théorique de la variable aléatoire Y
$\overline{\mathrm{X}}_n$	Moyenne empirique $\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i$ de l'échantillon $\mathbf{X}_n = (\mathrm{X}_1,\ldots,\mathrm{X}_n)^\intercal$, estimateur de μ_{X}
$\overline{x}_n$	Réalisation de la moyenne empirique $\frac{1}{n}\sum_{i=1}^n \mathrm{X}_i$ de l'échantillon $\mathbf{X}_n = (\mathrm{X}_1,\ldots,\mathrm{X}_n)^\intercal$, estimation de μ_{X}
$\xrightarrow{P}$	Symbole de convergence en probabilité
$\hat{\mathrm{F}}_n(\cdot) := \hat{\mathrm{F}}_{\mathbf{X}_n}(\cdot)$	Fonction de répartition empirique de l'échantillon $\mathbf{X}_n$
θ	Paramètre inconnu (parfois on notera $\theta^\bullet$ la vraie valeur inconnue du paramètre)
$\hat{\theta}(\mathrm{X}_1,\ldots,\mathrm{X}_n)$ ou $\hat{\theta}$	Estimateur du paramètre inconnu θ basé sur l'échantillon $\mathbf{X}_n = (\mathrm{X}_1,\ldots,\mathrm{X}_n)^\intercal$
$\hat{\theta}(x_1,\ldots,x_n)$ ou $\hat{\theta}$	Estimation du paramètre inconnu θ basé sur l'échantillon observé $\boldsymbol{x}_n = (x_1,\ldots,x_n)^\intercal$
$\mathbb{B}(\hat{\theta}(\mathrm{X}_1,\ldots,\mathrm{X}_n);\theta)$	Biais de l'estimateur $\hat{\theta}(\mathrm{X}_1,\ldots,\mathrm{X}_n)$ pour estimer le paramètre inconnu θ
$P[A]$	Probabilité de l'ensemble A
$\mathcal{V}(\theta;\mathrm{X}_1,\ldots,\mathrm{X}_n)$	Fonction de vraisemblance de l'échantillon $\mathbf{X}_n$ évaluée en θ
$\boldsymbol{x}^* = (x_1^*,\ldots,x_n^*)^\intercal$	Échantillon *bootstrap* généré à partir de l'échantillon observé $\boldsymbol{x}_n = (x_1,\ldots,x_n)^\intercal$
$\hat{\sigma}$	Estimateur de σ
$\hat{\sigma}$	Estimation de σ
p	Proportion théorique
$\hat{\mathrm{p}}$	Estimateur d'une proportion (ou d'une probabilité)
$\hat{p}$	Estimation d'une proportion (ou d'une probabilité)
$\widehat{\mathrm{m}_e}$	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 *bootstrap* 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 $\mathcal{F}(n,m)$
$\mathrm{IC}_{1-\alpha}(\theta)$	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
$(x_{(1)},\ldots,x_{(n)})$	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
$\hat{\mathrm{Y}}^p$	Préviseur de la variable aléatoire Y pour une nouvelle valeur de la variable explicative X en régression
$\mathrm{IP}_{1-\alpha}(\mathrm{Y}_0, x_0)$	Intervalle de prévision de niveau $1-\alpha$ pour la variable aléatoire Y_0 associée à une nouvelle valeur x_0 de la variable explicative
$\boldsymbol{\beta} = (\beta_0,\ldots,\beta_p)^\mathsf{T}$	Vecteur des $p+1$ coefficients inconnus d'un modèle de régression linéaire multiple à p variables explicatives

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\mathbf{Y}$	Estimateur du vecteur des paramètres inconnus $\boldsymbol{\beta}$ pour la matrice observée des variables explicatives $\boldsymbol{X}$ et pour le vecteur des valeurs à expliquer dans un modèle de régression linéaire multiple
$\hat{\boldsymbol{\beta}}$	Estimation de $\boldsymbol{\beta}$
VIF	Facteur d'inflation de la variance, *variance inflation factor*
AIC	*an information criterion*
BIC	*bayesian information criterion*
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
$\mu_{\bullet\bullet}$	Effet moyen général en ANOVA
$\mu_{i\bullet}$	Effet du niveau i d'un facteur en ANOVA
$\mu_{\bullet j}$	Effet du niveau j d'un facteur en ANOVA

Collection
Statistique
et probabilités
appliquées

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.

Pierre Lafaye de Micheaux,
Rémy Drouilhet, Benoît Liquet

Le logiciel R

Deuxième édition

Maîtriser le langage
Effectuer des analyses (bio)statistiques

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.

978-2-7462-4818-2

editions.lavoisier.fr