

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.

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) Initiation à la 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.