Avant-propos

Ce livre s'adresse à toute personne ayant à traiter des jeux de données, indépendamment du domaine d'application. Cette pratique impliquant typiquement de grandes quantités d'informations, l'aspect numérique est bien entendu primordial. De fait, il existe aujourd'hui de nombreux outils répondant à ces besoins. Nous avons opté ici pour le logiciel R dont le triple intérêt est d'être gratuit, très complet et en essor permanent. Néanmoins, aucune connaissance sur celui-ci n'est prérequise. Le livre se divise en effet en deux grandes parties : la première est centrée sur le logiciel lui-même, la seconde sur la mise en œuvre de méthodes statistiques classiques avec R.

Nous présentons les concepts de base du logiciel dans le premier chapitre. Le deuxième traite de la manipulation des données, c'est-à-dire des opérations courantes en statistique. Le bilan d'une étude passant par une visualisation claire des résultats, nous décrivons alors, en chapitre 3, certaines possibilités offertes par R dans ce domaine. Nous y présentons aussi bien la construction de graphiques simples que certaines variantes plus avancées. Les bases de la programmation sont quant à elles présentées au chapitre 4 : nous expliquons comment construire ses propres fonctions mais exposons aussi quelques-unes des procédures prédéfinies pour automatiser et paralléliser des analyses répétitives. Témoin du développement continu de R, le chapitre 5 se veut une introduction à quelques outils récents dédiés à des données en constante évolution, tant dans leurs formes que dans leur volume. Focalisée sur le logiciel R, cette première partie permet de comprendre les commandes apparaissant dans les méthodes exposées par la suite.

La seconde partie du livre propose de balayer un large spectre de techniques aussi bien classiques que récentes en traitement des données : intervalles de confiance et tests, procédures d'analyse factorielle, classification non supervisée, méthodes usuelles de régression, machine learning, gestion de données manquantes, analyse de texte, fouille de graphe, etc. Chaque méthode est illustrée sur un exemple et traitée de façon autonome dans une fiche spécifique. Après une brève présentation du contexte, les lignes de commandes R sont détaillées et les résultats commentés.

On pourra télécharger les jeux de données et retrouver tous les résultats décrits ainsi que les solutions des exercices proposés en première partie sur le site du livre : `https://r-stat-sc-donnees.github.io/`