Introduction

Une bonne partie des actions effectuées à partir d'un logiciel statistique revient à manipuler, voire transformer des données numériques représentant des données statistiques au sens propre. Il est donc primordial de bien comprendre comment sont représentées les données statistiques et comment elles peuvent être exploitées par un logiciel tel que R. Après l'importation, le recodage et la transformation éventuelle de ces données, la description des variables d'intérêt et le résumé de leur distribution sous forme numérique et graphique constitue une étape préalable et fondamentale à toute modélisation statistique, d'où l'importance de ces premières étapes dans le déroulement d'un projet d'analyse statistique. Dans un second temps, il est essentiel de bien maîtriser les commandes permettant de calculer les principales mesures d'association en recherche médicale et de savoir mettre en œuvre les modèles explicatifs et prédictifs classiques : analyse de la variance, régression linéaire et logistique et modèle de Cox. A quelques exceptions près, on préfèrera recourir aux commandes R disponibles lors de l'installation du logiciel (commandes de base), plutôt qu'à des librairies spécialisées de commandes, encore appelées packages sous R. Les packages à installer pour suivre les applications proposées dans cet ouvrage sont listés dans le premier chapitre, à la section « Avant de démarrer ».

Ce livre suppose que le lecteur est déjà familier avec les notions statistiques de base, en particulier le calcul des indicateurs de tendance centrale et de dispersion pour une variable continue, les tableaux de contingence, l'analyse de la variance et les modèles de régression classique. L'objectif est ici de mettre en application ces connaissances sur des jeux de données décrits dans de nombreux autres ouvrages, même si l'interprétation des résultats reste minimale, afin de se familiariser rapidement avec l'usage de R sur des données réelles. En particulier, l'accent est mis sur la gestion et la manipulation de données structurées puisqu'il s'avère que cela constitue 60 à 80 % du travail du statisticien. Il existe de nombreux ouvrages en français ou en anglais sur R, tant du point de vue technique que statistique. Certains de ces ouvrages sont plutôt à caractère général [SHA 12], d'autres au contraire sont

beaucoup plus spécialisés [BIL 14] ou abordent des notions plus avancées [HOT 09]. Le but de ce livre est de permettre au lecteur de se familiariser rapidement avec R afin qu'il puisse réaliser ses propres analyses et continuer son apprentissage de manière autonome dans le domaine des statistiques médicales.

Dans le premier chapitre, on introduira les commandes de base pour la gestion de données sous R. Il s'agit principalement de la création et de la manipulation de variables quantitatives et qualitatives (recodage de valeurs individuelles, comptage des observations manquantes), de l'importation de bases de données stockées sous forme de fichiers texte, ainsi que d'opérations arithmétiques élémentaires (minimum, maximum, moyenne arithmétique, différence, fréquence, etc.). On verra également comment sauvegarder des bases de données prétraitées au format texte ou R. L'objectif est de comprendre comment les données sont représentées sous R, et comment travailler à partir de celles-ci.

Le deuxième chapitre porte sur les commandes utiles pour la description d'un tableau de données constitué de variables quantitatives ou qualitatives. L'approche descriptive est strictement univariée, ce qui constitue le préalable à toute démarche statistique. Les commandes graphiques de base (histogramme, courbe de densité, diagramme en barres ou en points) seront présentées en complément des résumés descriptifs numériques usuels de tendance centrale (moyenne, médiane) et de dispersion (variance, quartiles). On abordera également l'estimation ponctuelle et par intervalle à l'aide d'une moyenne arithmétique et d'une proportion empirique. L'objectif est de se familiariser avec l'emploi de commandes R simples opérant sur une variable, éventuellement en précisant certaines options pour le calcul, et à la sélection d'unités statistiques parmi l'ensemble des observations disponibles.

Le troisième chapitre est consacré à la comparaison de deux échantillons, pour des mesures quantitatives ou qualitatives. On aborde les tests d'hypothèse suivants : test de Student pour échantillons indépendants ou appariés, test non paramétrique de Wilcoxon, test du χ^2 et test exact de Fisher, test de NcNemar, à partir des principales mesures d'association pour deux variables (différence de moyennes, odds-ratio et risque relatif). A partir de ce chapitre, on insistera moins sur la description univariée de chaque variable, mais il est conseillé de toujours procéder aux étapes de description des données abordées dans le deuxième chapitre. L'objectif est de maîtriser les principaux tests statistiques dans le cas où l'on s'intéresse à la relation entre une variable quantitative et une variable qualitative, ou pour deux variables qualitatives.

Le quatrième chapitre est une introduction à l'analyse de variance dans laquelle on cherche à expliquer la variabilité observée au niveau d'une variable réponse numérique par la prise en compte d'un facteur de groupe ou de classification et à l'estimation par intervalle de différences de moyennes. On mettra l'accent sur la construction d'un tableau d'ANOVA résumant les différentes sources de variabilité et sur les méthodes graphiques permettant de résumer la distribution des données individuelles

ou aggrégées. On discutera également le test de tendance linéaire lorsque le facteur de classification peut être considéré comme naturellement ordonné. L'objectif est de comprendre comment construire un modèle explicatif dans le cas où l'on a un, voire deux, facteurs explicatifs, et comment présenter numériquement et graphiquement les résultats d'un tel modèle à l'aide de R.

Le cinquième chapitre porte sur l'analyse de la relation linéaire entre deux variables quantitatives continues. Dans l'approche de corrélation linéaire, qui suppose une relation symétrique entre les deux variables, on s'intéressera à quantifier la force et la direction de l'association de manière paramétrique (corrélation de Pearson) ou non paramétrique (corrélation de Spearman basée sur les rangs) et à la représentation graphique de cette relation. La régression linéaire simple sera utilisée dans le cas où l'une des deux variables numériques joue le rôle d'une variable réponse et l'autre celui d'une variable explicative. On présentera les commandes utiles pour l'estimation des coefficients de la droite de régression, la construction du tableau d'ANOVA associé à la régression, et la prédiction. L'objectif de ce chapitre reste identique à celui du quatrième chapitre, à savoir présenter les commandes R nécessaires à la construction d'un modèle statistique simple entre deux variables, dans une optique explicative ou prédictive.

Dans le sixième chapitre seront abordées les principales mesures d'association rencontrées dans les études épidémiologiques : odds-ratio, risque relatif, prévalence, etc. Les commandes R permettant l'estimation (ponctuelle et par intervalle) et les tests d'hypothèse associés seront illustrées sur des données de cohorte ou d'études cas-témoins. La mise en œuvre d'un modèle de régression logistique simple permet de compléter l'éventail des méthodes statistiques permettant d'expliquer la variabilité observée au niveau d'une variable réponse binaire. L'objectif est de comprendre les commandes R à utiliser dans le cas où les variables sont binaires, soit pour résumer un tableau de contingence sous forme d'indicateurs d'association soit pour modéliser la relation entre une réponse binaire (malade/non-malade) et une variable explicative qualitative à partir de données dites groupées.

Le septième et dernier chapitre constitue une introduction à l'analyse de données censurées, aux principaux tests associés à la construction d'une courbe de survie (test du log-rank ou de Wilcoxon) et enfin au modèle de régression de Cox. La spécificité des données censurées impose un soin particulier dans le codage des données sous R, et l'objectif est de présenter les commandes R essentielles à la bonne représentation des données de survie sous forme numérique, à leur résumé numérique (médiane de survie) et graphique (courbe de Kaplan-Meier), et à la mise en œuvre des tests courants.

A la fin de chaque chapitre, quelques applications sont proposées et des exemples de commandes permettant de répondre aux questions posées sont proposées pour la plupart des questions. Il est parfois possible d'obtenir des résultats identiques

par d'autres approches ou en utilisant d'autres commandes. Les sorties R ne sont pas reproduites mais le lecteur est encouragé à essayer lui-même les instructions R proposées et à essayer des instructions alternatives ou complémentaires. On supposera que les fichiers de données utilisés sont disponibles dans le répertoire de travail. L'ensemble des fichiers de données et les commandes R utilisées dans cet ouvrage peuvent être téléchargés sur le site GitHub, à l'adresse https://github.com/biostatsante.

Trois annexes permettent de mieux se familiariser avec RStudio, les packages `lattice` pour la gestion des sorties graphiques et `Hmisc` et `rms` pour la gestion de données avancée et la modélisation. Ces chapitres ne remplacent évidemment pas les excellents ouvrages de John Verzani, Paul Murrell et Franck Harrell [VER 11, MUR 05, HAR 01].

Pour des raisons de mise en page, certaines sorties R ont été tronquées ou reformatées. Par conséquent, celles-ci pourraient être amenées à différer lorsque le lecteur tente de reproduire les commandes contenues dans cet ouvrage.

Un index des commandes R utilisées dans les illustrations est disponible à la fin de l'ouvrage.