

BIO4558 Biostatistiques appliquées avec R

Manuel de Laboratoire

Julien Martin

2019-09-25

Contents

Préface	5
Quelques points importants à retenir	5
Qu'est-ce que R et pourquoi l'utiliser dans ce cours?	6
Installation	7
Instructions générales pour les laboratoires	7
Et pour les dilettantes ou ceux qui sont allergiques à la programmation...	8

Préface

Les exercices de laboratoire que vous retrouverez dans les pages qui suivent sont conçus de manière à vous permettre de développer une expérience pratique en analyse de données à l'aide d'un logiciel (R). R est un logiciel très puissant, mais comme tous les logiciels, il a des limites. En particulier il ne peut réfléchir à votre place, vous dire si l'analyse que vous tentez d'effectuer est appropriée ou sensée, ou interpréter biologiquement les résultats.

Quelques points importants à retenir

- Avant de commencer une analyse statistique, il faut d'abord vous familiariser son fonctionnement. Cela ne veut pas dire que vous devez connaître les outils mathématiques qui la sous-tendent, mais vous devriez au moins comprendre les principes utilisés lors de cette analyse. Avant de faire un exercice de laboratoire, lisez donc la section correspondante dans les notes de cours. Sans cette lecture préalable, il est très probable que les résultats produits par le logiciel, même si l'analyse a été effectuée correctement, seront indéchiffrables.
- Les laboratoires sont conçus pour compléter les cours théoriques et vice versa. À cause des contraintes d'horaires, il se pourrait que le cours et le laboratoire ne soient pas parfaitement synchronisés. N'hésitez donc pas à poser des questions sur le labo en classe ou des questions théoriques au laboratoire.
- Travaillez sur les exercices de laboratoire à votre propre rythme. Certains exercices prennent beaucoup moins de temps que d'autres et il n'est pas nécessaire de compléter un exercice par séance de laboratoire. En fait deux séances de laboratoire sont prévues pour certains des exercices. Même si vous n'êtes pas notés sur les exercices de laboratoire, soyez conscient que ces exercices sont essentiels. Si vous ne les faites pas, il est très peu probable que vous serez capable de compléter les devoirs et l'examen final. Prenez donc ces exercices de laboratoire au sérieux !

- Le premier laboratoire est conçu pour vous permettre d'acquérir ou de réviser le minimum de connaissances requises pour vous permettre de réaliser les exercices de laboratoires avec R. Il y a presque toujours de multiples façons de faire les choses avec R et vous ne trouverez ici que des méthodes simples. Ceux et celles d'entre vous qui y sont enclins pourront trouver en ligne des instructions plus détaillées et complexes. En particulier, je vous conseille :
 - R pour les débutants http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf
 - Using R for psychological research: A simple guide to an elegant package <http://www.personality-project.org/r/>
 - An introduction to R <http://cran.r-project.org/doc/manuals/R-intro.html>
 - Si vous préférez des manuels, le site web de CRAN en garde une liste commentée à : <http://www.r-project.org/doc/bib/R-books.html>
 - Finalement, comme aide-mémoire à garder sous la main, je vous recommande R reference card par Tom Short <http://cran.r-project.org/doc/contrib/Short-refcard.pdf>

Qu'est-ce que R et pourquoi l'utiliser dans ce cours?

R est un logiciel libre et multiplateforme formant un système statistique et graphique. R est également un langage de programmation spécialisé pour les statistiques. C'est un dialecte du langage S. S-Plus est un autre dialecte, très semblable, et forme un produit commercial qui a un interface graphique que certains trouvent plus convivial.

R a deux très grands avantages pour ce cours, et un inconvénient embêtant initialement mais qui vous forcera à acquérir des excellentes habitudes de travail. Le premier avantage est que vous pouvez tous l'installer sur votre (ou vos) ordinateurs personnel gratuitement. C'est important parce que c'est à l'usage que vous apprendrez et maîtriserez réellement les biostatistiques et cela implique que vous devez avoir un accès facile et illimité à un logiciel statistique. Le deuxième avantage est que R peut tout faire en statistiques. R est conçu pour être extensible et est devenu l'outil de prédilection des statisticiens mondialement. La question n'est plus : " Est-ce que R peut faire ceci? ", mais devient " Comment faire ceci avec R ". Et la recherche internet est votre ami. Aucun autre logiciel n'offre ces deux avantages.

L'inconvénient embêtant initialement est que l'on doit opérer R en tapant des instructions (ou en copiant des sections de code) plutôt qu'en utilisant des menus et en cliquant sur différentes options. Si on ne sait pas quelle commande taper, rien ne se passe. Ce n'est donc pas facile d'utilisation à priori. Cependant, il

est possible d'apprendre rapidement à faire certaines des opérations de base (ouvrir un fichier de données, faire un graphique pour examiner ces données, effectuer un test statistique simple). Et une fois que l'on comprend le principe de la chose, on peut assez facilement trouver sur le web des exemples d'analyses ou de graphiques plus complexes et adapter le code à nos propres besoins. C'est ce que vous ferez dans le premier laboratoire pour vous familiariser avec R.

Pourquoi cet inconvénient est-il d'une certaine façon un avantage? Parce que vous allez sauver du temps en fin de compte. Garanti. Croyez-moi, on ne fait jamais une analyse une seule fois. En cours de route, on découvre des erreurs d'entrée de données, ou que l'on doit faire l'analyse séparément pour des sous-groupes, ou on obtient des données supplémentaires, ou on fait une erreur. On doit alors recommencer l'analyse. Avec une interface graphique et des menus, cela implique recommencer à cliquer ici, entre des paramètres dans des boîtes et sélectionner des boutons. Chaque fois avec possibilité d'erreur. Avec une série de commandes écrites, il suffit de corriger ce qui doit l'être puis de copier-coller l'ensemble pour répéter instantanément. Et vous avez la possibilité de parfaitement documenter ce que vous avez fait. C'est comme cela que les professionnels travaillent et offrent une assurance de qualité de leurs résultats.

Installation

Pour installer R sur un nouvel ordinateur, allez au site <http://cran.r-project.org/>. Vous y trouverez des versions compilées (binaries) ou non (sources) pour votre système d'exploitation de prédilection (Windows, MacOS, Linux).

Note : R a déjà été installé sur les ordinateurs du laboratoire (la version pourrait être un peu plus ancienne, mais cela devrait être sans conséquences).

Instructions générales pour les laboratoires

- Apporter une clé USB ou son équivalent à chaque séance de laboratoire pour sauvegarder votre travail.
- Lire l'exercice de laboratoire AVANT la séance, lire le code R correspondant et préparer vos questions sur le code.
- Durant les pré-labs, écouter les instructions et posez vos questions au moment approprié.
- Faites les exercices du manuel de laboratoire à votre rythme, en équipe, puis je vous recommande de commencer (compléter?) le devoir. Profitez de la présence du démonstrateur et du prof...
- Pendant vos analyses, copiez-collez des fragments de sorties de R dans un document (par exemple dans votre traitement de texte favori). Annotez abondamment.

- À chaque fois que vous fermez R, sauvegardez l'historique de vos commandes (ex: labo1.1.rHistory, labo1.2.rHistory, etc). Vous pourrez ainsi refaire le labo instantanément, récupérer des fragments de code, ou plus facilement identifier les erreurs dans vos analyses.
- Créez votre propre librairie de fragments de codes (snippets). Annotez-là abondamment. Vous vous en félicitez plus tard.

Et pour les dilettantes ou ceux qui sont allergiques à la programmation...

Je vous conseille d'essayer d'appivoiser R sans interface graphique. Au début ce sera peut-être difficile, mais cela vous forcera à acquérir de bonnes habitudes. Si vous trouvez cela vraiment trop lourd, ou que vous vous heurtez à des problèmes récurrents qui vous font perdre trop de temps, alors vous pouvez installer (du moins sur votre propre ordi, je ne sais pas si vous pouvez au laboratoire) une interface graphique distribuée par CRAN et qui fonctionne sous Windows, Mac et Linux (à ce qu'il paraît, je n'ai essayé que la version Windows). Il s'agit de Rcmdr, que vous devez d'abord installer dans R. Vous devez avoir accès à l'Internet. Dans la fenêtre de commande, entrez la commande

```
install.packages("Rcmdr", dependencies=TRUE)
```

Une fenêtre s'ouvrira vous permettant de choisir l'un des sites distribuant Rcmdr. Le téléchargement et l'installation prendront plusieurs minutes car plusieurs autres packages doivent être téléchargés et installés. Quand ce sera terminé, vous pourrez démarrer l'interface graphique en entrant la commande: `library(Rcmdr)` Une nouvelle fenêtre s'ouvrira, avec des menus pour accéder à la majorité des commandes usuelles. Une des sous-fenêtres de Rcmdr est un historique des commandes R correspondant à vos choix dans les menus. Étudiez-les pour apprendre, peut-être plus facilement, le langage R.