

TP de Biométrie Semestre 3

Benoît Simon-Bouhet

2022-09-16T00:00:00+02:00

Table des matières

Introduction

Objectifs

Ce livre contient l'ensemble du matériel (contenus, exemples, exercices...) nécessaire à la réalisation des travaux pratiques de **Biométrie** de l'EC '*Outils pour l'étude et la compréhension du vivant 2*' du semestre 3 de la licence Sciences de la Vie de La Rochelle Université.

Les 4 grands chapitres de ce livre correspondent aux 3 objectifs principaux de ces séances de TP et TEA :

1. **Vous faire découvrir les logiciels R et Rstudio** (chap. 1 et 2) dans lesquels vous allez passer beaucoup de temps en licence puis en master. Si vous choisissez une spécialité de master qui implique de traiter des données (c'est-à-dire à peu près toutes les spécialités des Sciences de la Vie !) et/ou de communiquer des résultats d'analyses statistiques, alors R et RStudio devraient être les logiciels vers lesquels vous vous tournerez naturellement.
2. **Vous apprendre à faire des graphiques de qualités dans RStudio et vous faire prendre conscience de l'importance des visualisations graphiques** (chap. 3) :
 - d'une part, pour explorer des données inconnues et vous faire une première idée des informations qu'elles contiennent,
 - d'autre part, pour vous permettre de formuler des hypothèses pertinentes et intéressantes concernant les systèmes que vous étudiez,
 - et enfin, pour communiquer efficacement vos trouvailles à un public qui ne connaît pas vos données aussi bien que vous (cela inclut évidemment vos enseignants à l'issue de vos stages).
3. **Vous apprendre à manipuler efficacement des tableaux de données de grande taille** (chap. 4). Cela signifie que vous devriez être mesure de sélectionner des variables (colonnes) d'un tableau, d'en créer de nouvelles en modifiant et/ou combinant des variables existantes, de filtrer des lignes spécifiques, etc.

À l'issue de ces TP et TEA, vous devriez donc être suffisamment à l'aise avec le logiciel **RStudio** pour y importer des données issues de tableurs, les manipuler pour les mettre dans un format permettant les représentations graphiques, et pour produire des graphiques pertinents, adaptés aux données dont vous disposez, et d'une qualité vous permettant de les intégrer sans honte à vos compte-rendus de TP et rapports de stages.

D'ailleurs, les données que vous serez amenés à traiter lors de vos stages, ou plus tard, lorsque vous serez en poste, ont souvent été acquises à grands frais, et au prix d'efforts importants. Il est donc de votre responsabilité d'en tirer le maximum. Et ça commence toujours (ou presque), par la manipulation de données dans **RStudio** et réalisation de visualisations graphiques parlantes.

Dernière choses : à partir de maintenant, tous les compte-rendus de TP que vous aurez à produire dans le cadre de la licence SV devront respecter les bonnes pratiques décrites dans ce document. En particulier, les collègues de l'équipe pédagogique attendent en effet que les graphiques que vous intégrerez à vos compte-rendus de TP soient systématiquement produits dans **RStudio**.

Organisation

Volume de travail

Les travaux pratiques et TEA de biométrie auront lieu entre le 12 septembre et le 07 octobre 2022 :

- Semaine 37 (du 12 au 16 septembre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 38 (du 19 au 23 septembre) : 1 séance de TP d'1h30 et 1 séance de TEA d'1h30
- Semaine 39 (du 26 au 30 septembre) : 1 séance de TP d'1h30
- Semaine 40 (du 03 au 07 octobre) : 1 séance de TP d'1h30

Tous les TP ont lieu en salle MSI 217. Tous les TEA sont à distance.

Au total, chaque groupe aura donc 4 séances de TP et 2 séances de TEA. C'est très peu pour atteindre les objectifs fixés et il y aura donc évidemment du travail personnel à fournir en dehors de ces séances. Pour chaque TP ou TEA prévu à l'emploi du temps, j'estime qu'une à deux heures de travail personnel est nécessaire (soit 9 à 15 heures au total, selon votre degré d'aisance, à répartir sur 4 semaines). Attention donc : pensez bien à prévoir du temps dans vos plannings car le travail personnel est essentiel pour progresser dans cette matière. J'insiste sur l'importance de faire l'effort dès maintenant : vous allez en effet avoir des enseignements qui reposent sur l'utilisation de ces logiciels à chaque semestre de la licence du S3 au S6. C'est maintenant qu'il faut acquérir des automatismes, cela vous fera gagner énormément de temps ensuite.

Modalités d'enseignement

Pour suivre cet enseignement vous pourrez utiliser les ordinateurs de l'université, mais je ne peux que vous encourager à utiliser vos propres ordinateurs, sous Windows, Linux ou MacOS. Lors de vos futurs stages et pour rédiger vos comptes-rendus de TP, vous utiliserez le plus souvent vos propres ordinateurs, autant prendre dès maintenant de bonnes habitudes en installant les logiciels dont vous aurez besoin tout au long de votre licence. Si vous ne possédez pas d'ordinateur, manifestez vous rapidement auprès de moi car des solutions existent (prêt par l'université, travail sur tablette via [RStudio cloud...](#)).

! Important

L'essentiel du contenu de cet enseignement peut être abordé en autonomie, à distance, grâce à ce livre en ligne, aux ressources mises à disposition sur Moodle et à votre ordinateur personnel. Cela signifie que **la présence physique lors de ces séances de TP n'est pas obligatoire.**

Plus que des séances de TP classiques, considérez plutôt qu'il s'agit de **permanences non-obligatoires** : si vous pensez avoir besoin d'aide, si vous avez des points de blocage ou des questions sur le contenu de ce document, sur les exercices demandés ou sur les quizz Moodle, alors venez poser vos questions lors des séances de TP. Vous ne serez d'ailleurs pas tenus de rester pendant 1h30 : si vous obtenez une réponse en 10 minutes et que vous préférez travailler ailleurs, vous serez libres de repartir !

De même, si vous n'avez pas de difficulté de compréhension, que vous n'avez pas de problème avec les exercices de ce livre en ligne ni avec les quizz Moodle, votre présence n'est pas requise. Si vous souhaitez malgré tout venir en salle de TP, pas de problème, vous y serez toujours les bienvenus.

Ce fonctionnement très souple a de nombreux avantages :

- vous vous organisez comme vous le souhaitez
- vous ne venez que lorsque vous en avez vraiment besoin
- celles et ceux qui se déplacent reçoivent une aide personnalisée
- vous travaillez sur vos ordinateurs
- les effectifs étant réduits, c'est aussi plus confortable pour moi !

Toutefois, pour que cette organisation fonctionne, cela demande de la rigueur de votre part, en particulier sur la régularité du travail que vous devez fournir. Si la présence en salle de TP n'est pas requise, le travail demandé est bel et bien obligatoire ! Si vous venez en salle de TP sans avoir travaillé en amont, votre venue sera totalement inutile puisque vous n'aurez pas de question à poser et que vous passerez votre séance à lire ce livre en ligne. Vous perdrez donc votre temps, celui de vos collègues, et le mien. De même, si vous attendez la 4e semaine pour vous y mettre, vous irez droit dans le mur. Je le répète, outre les 9h de TP/TEA prévus

dans vos emplois du temps, vous devez prévoir entre 9 et 15 heures de travail personnel supplémentaire.

Je vous laisse donc une grande liberté d'organisation. À vous d'en tirer le maximum et de faire preuve du sérieux nécessaire. Le rythme auquel vous devriez avancer est présenté dans la partie suivante intitulée "Progression conseillée".

Utilisation de Slack

Outre les séances de permanence non-obligatoires, nous échangerons aussi sur [l'application Slack](#), qui fonctionne un peu comme un "twitter privé". Slack facilite la communication des équipes et permet de travailler ensemble. Créez-vous un compte en ligne et installez le logiciel sur votre ordinateur (il existe aussi des versions pour tablettes et smartphones). Lorsque vous aurez installé le logiciel, [cliquez sur ce lien](#) pour vous connecter à notre espace de travail commun intitulé **L2 SV 22-23 / EC outils** (ce lien expire régulièrement : faites moi signe s'il n'est plus valide).

Vous verrez que 2 "chaînes" sont disponibles :

- #général : c'est là que les questions liées à l'organisation générale du cours, des TP et TEA doivent être posées. Si vous ne savez pas si une séance de permanence a lieu, posez la question ici.
- #questions-rstudio : c'est ici que toutes les questions pratiques liées à l'utilisation de R et RStudio devront être posées. Problèmes de syntaxe, problèmes liés à l'interface, à l'installation des packages ou à l'utilisation des fonctions, à la création des graphiques, à la manipulation des tableaux... Tout ce qui concerne directement les logiciels sera traité ici. Vous êtes libres de poser des questions, de poster des captures d'écran, des morceaux de code, des messages d'erreur. Et **vous êtes bien entendus vivement encouragés à vous entraider et à répondre aux questions de vos collègues**. Je n'interviendrai ici que pour répondre aux questions laissées sans réponse ou si les réponses apportées sont inexactes. Le fonctionnement est celui d'un forum de discussion instantané. Vous en tirerez le plus grand bénéfice en participant et en n'ayant pas peur de poser des questions, même si elles vous paraissent idiotes. Rappelez-vous toujours que si vous vous posez une question, d'autres se la posent aussi probablement.

Ainsi, quand vous travaillerez à vos TP ou TEA, que vous soyez installés chez vous ou en salle de TP, prenez l'habitude de garder Slack ouvert sur votre ordinateur. Même si vous n'avez pas de question à poser, votre participation active pour répondre à vos collègues est souhaitable et souhaitée. Je vous incite donc fortement à vous **entraider** : c'est très formateur pour celui qui explique, et celui qui rencontre une difficulté a plus de chances de comprendre si c'est quelqu'un d'autre qui lui explique plutôt que la personne qui a rédigé les instructions mal comprises.

Ce document est fait pour vous permettre d'avancer en autonomie et vous ne devriez normalement pas avoir beaucoup besoin de moi si votre lecture est attentive. L'expérience montre en effet que la plupart du temps, il suffit de lire correctement les paragraphes précédents et/ou suivants pour obtenir la réponse à ses questions. J'essaie néanmoins de rester disponible sur Slack pendant les séances de TP et de TEA de tous les groupes. Cela veut donc dire que même si votre groupe n'est pas en TP, vos questions ont des chances d'être lues et de recevoir des réponses dès que d'autres groupes sont en TP ou TEA. Vous êtes d'ailleurs encouragés à échanger sur Slack aussi pendant vos phases de travail personnels.

Progression conseillée

Pour apprendre à utiliser un logiciel comme R, il faut faire les choses soi-même, ne pas avoir peur des messages d'erreurs (il faut d'ailleurs apprendre à les déchiffrer pour comprendre d'où viennent les problèmes), essayer maintes fois, se tromper beaucoup, recommencer, et surtout, ne pas se décourager. J'utilise ce logiciel presque quotidiennement depuis plus de 15 ans et à chaque session de travail, je rencontre des messages d'erreur. Avec suffisamment d'habitude, on apprend à les déchiffrer, et on corrige les problèmes en quelques secondes. Ce livre est conçu pour vous faciliter la tâche, mais ne vous y trompez pas, vous rencontrerez des difficultés, et c'est normal. C'est le prix à payer pour profiter de la puissance du meilleur logiciel permettant d'analyser des données, de produire des graphiques de qualité et de réaliser toutes les statistiques dont vous aurez besoin d'ici la fin de vos études et au-delà.

Pour que cet apprentissage soit le moins problématique possible, il convient de prendre les choses dans l'ordre. C'est la raison pour laquelle les chapitres de ce livre doivent être lus dans l'ordre, et les exercices d'application faits au fur et à mesure de la lecture.

Idéalement, voilà les étapes que vous devriez avoir franchi chaque semaine :

1. La première semaine (37) est consacrée à l'installation des logiciels, à la découverte de l'environnement de travail, des **RProjects**, des packages et des scripts. Avant votre deuxième séance, vous devrez être capables de créer un Rproject et un script, de télécharger et d'installer des packages, et d'exécuter des commandes simples dans votre script. Vous devrez avoir suivi les tutoriels de DataCamp et connaître les types d'objets suivants : vecteurs, facteurs, data.frames (et tibbles). Vous devrez effectuer le premier quizz Moodle (attention, ce quizz est à faire seul !).
2. La deuxième semaine (38) est consacrée à la découverte des premiers jeux de données et du package **ggplot2**. Avant votre troisième séance, vous devrez avoir compris comment charger en mémoire les jeux de données disponibles dans un packages, vous devrez connaître la syntaxe de base permettant de faire toutes sortes de graphiques avec **ggplot2** avec une ou deux variables. À terme, vous devrez être capables de choisir des graphiques appropriés selon le nombre et la nature des variables dont vous disposez. À ce stade, on ne demande rien de complexe, mais vous devrez, à minima, être capable de faire des

barplots, des histogrammes et des nuages de points. Vous devrez effectuer le second quizz Moodle (attention, ce quizz est à faire seul !).

3. La troisième semaine (39) est consacrée à l'amélioration de la qualité de vos graphiques. Avant votre dernière séance de TP, vous devrez être capable de faire, outre les graphiques de la semaine précédente, des stripcharts, et des graphiques en lignes et des boxplots. Vous devrez également être capable de faire des sous-graphiques par catégories (`facet()`), de choisir un thème et des palettes de couleurs appropriées, et de légender/annoter correctement vos graphiques. Vous devrez effectuer le troisième quizz Moodle (attention, ce quizz est à faire seul !).
4. La quatrième semaine (40) est consacrée à la manipulation des tableaux de données. Avant la fin de cette semaine, vous devrez être capable de sélectionner des colonnes dans un tableau, de filtrer des lignes, et de créer de nouvelles variables. Vous devrez être capables d'enchaîner correctement les étapes suivantes : ouvrir le logiciel > créer un projet > créer un script > mettre en mémoire les packages utiles > importer des données > mettre en forme ces données > faire un ou des graphiques informatifs et correctement mis en forme. Vous devrez effectuer le quatrième et dernier quizz Moodle (attention, ce quizz est à faire seul !).

Évaluation(s)

Vous serez évalués à 3 niveaux :

1. Votre participation sur Slack
2. Les quizz Moodle
3. Une évaluation plus classique

Les exercices demandés dans ce document en ligne ne seront ni ramassés, ni notés : ils sont proposés pour que vous puissiez mettre en application les notions récemment apprises et afin d'évaluer votre propre progression et vos apprentissages. Mais tout ce que nous voyons en TP et TEA devra être acquis à la fin des TP.

En particulier, je vérifierai que les étapes décrites précédemment sont maîtrisées (ouvrir le logiciel > créer un projet > créer un script > mettre en mémoire les packages utiles > importer des données > mettre en forme ces données > faire un ou des graphiques informatifs et correctement mis en forme) en vous fournissant un jeu de données inconnu que vous devrez importer et mettre en forme dans **RStudio** afin de produire quelques graphiques informatifs. Cette évaluation aura lieu soit en salle informatique, soit dans le cadre d'un compte-rendu de TP que vous devrez rendre pour un collègue. Si cette modalité est mise en place, je vous en dirai plus en temps utile.

1 R et RStudio : les bases

1.1 Préambule

Avant de commencer à explorer des données dans R, il y a plusieurs concepts clés qu'il faut comprendre en premier lieu :

1. Que sont R et RStudio ?
2. Comment s'y prend-on pour coder dans R ?
3. Que sont les **packages** ?

Une bonne maîtrise des éléments présentés dans ce chapitre est indispensable pour aborder sereinement les chapitres suivants, à commencer par le chapitre Chapitre ??, qui présente un jeu de données que nous explorerons en détail un peu plus tard. Lisez donc attentivement ce chapitre et faites bien tous les exercices demandés.

Ce chapitre est en grande partie basé sur les 3 ressources suivantes que je vous encourage à consulter si vous souhaitez obtenir plus de détails :

1. L'ouvrage intitulé [ModernDive](#), de Chester Ismay et Albert Y. Kim. Une bonne partie de ce livre est très largement inspirée de cet ouvrage. C'est en anglais, mais c'est un très bon texte d'introduction aux statistiques sous R et RStudio.
2. L'ouvrage intitulé [Getting used to R, RStudio, and R Markdown](#) de Chester Ismay, comprend des podcasts (en anglais toujours) que vous pouvez suivre en apprenant.
3. Les tutoriels en ligne de [DataCamp](#). DataCamp est une plateforme de e-learning accessible depuis n'importe quel navigateur internet et dont la priorité est l'enseignement des "data sciences". Leurs tutoriels vous aideront à apprendre certains des concepts de développés dans ce livre.

! Important

Avant d'aller plus loin, rendez-vous sur [le site de DataCamp](#) et créez-vous un compte gratuit.

1.2 Que sont R et RStudio ?

Pour l'ensemble de ces TP, j'attends de vous que vous utilisiez **R** *via* **RStudio**. Les utilisateurs novices confondent souvent les deux. Pour tenter une analogie simple :

- **R** est le moteur d'une voiture
- **RStudio** est l'habitacle, le tableau de bord, les pédales...

Si vous n'avez pas de moteur, vous n'irez nulle part. En revanche, un moteur sans tableau de bord est difficile à manœuvrer. Il est en effet beaucoup plus simple de faire avancer une voiture depuis l'habitacle, plutôt qu'en actionnant à la main les câbles et leviers du moteur.

En l'occurrence, **R** est un langage de programmation capable de produire des graphiques et de réaliser des analyses statistiques, des plus simples aux plus complexes. **RStudio** est un "emballage" qui rend l'utilisation de **R** plus aisée. **RStudio** est ce qu'on appelle un IDE ou "Integrated Development Environment". On peut utiliser **R** sans **RStudio**, mais c'est nettement plus compliqué, nettement moins pratique.

1.2.1 Installation

Avertissement

Si vous travaillez exclusivement sur les ordinateurs de l'Université, vous pouvez passer cette section. En revanche, si vous souhaitez utiliser **R** et **RStudio** sur votre ordinateur personnel, alors lisez attentivement la suite !

Avant tout, vous devez télécharger et installer **R**, **puis** **RStudio**, dans cet ordre :

1. [Téléchargez et installez R](#)



- Vous devez installer ce logiciel en premier.
- Cliquez sur le lien de téléchargement qui correspond à votre système d'exploitation, puis, sur "base", si vous êtes sous Windows, sur "**R-4.2.1.pkg**" si vous êtes sous Mac avec processeur Intel, ou sur **R-4.2.1-arm64.pkg** si vous êtes sous Mac avec processeur M1 ou M2 (sous Mac, cliquez sur le Menu , puis sur "À propos de ce Mac" et regardez à la rubrique "Processeur), et suivez les instructions.

2. [Téléchargez et installez RStudio](#)

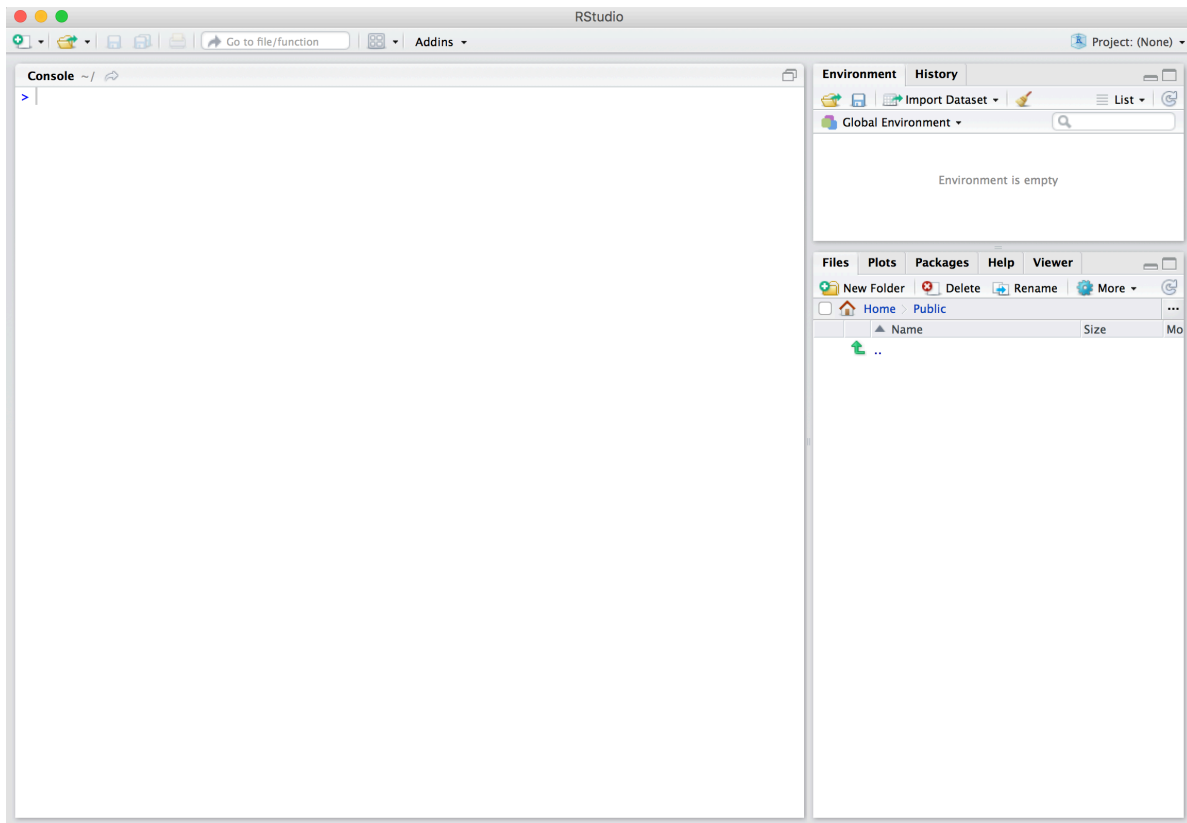
- Cliquez sur "RStudio Desktop", puis sur "Download RStudio Desktop".
- Choisissez la version gratuite et cliquez sur le lien de téléchargement qui correspond à votre système d'exploitation.

1.2.2 Utiliser R depuis RStudio

Puisqu'il est beaucoup plus facile d'utiliser Rstudio pour interagir avec R, nous utiliserons exclusivement l'interface de RStudio. Après les installations réalisées à la Section ??, vous disposez de 2 nouveaux logiciels sur votre ordinateur. RStudio ne peut fonctionner sans R, mais nous travaillerons exclusivement dans RStudio :

- R, ne pas ouvrir ceci : 
- RStudio, ouvrir cela : 

À l'université, vous trouverez RStudio dans le menu Windows. Quand vous ouvrez RStudio pour la première fois, vous devriez obtenir une fenêtre qui ressemble à ceci :



Prenez le temps d'explorer cette interface, cliquez sur les différents onglets, ouvrez les menus, allez faire un tour dans les préférences du logiciel pour découvrir les différents panneaux de l'application, en particulier la Console dans laquelle nous exécuterons très bientôt du code R.