Introduction

Welcome!

In this module of the course, we will be learning how to use R to analyze data. This workbook will contain pre-class readings, in-class materials, and links to additional resources.

Most of this workbook has been adapted from The Epidemiologist R Handbook. You can find more details at the bottom of this page.

Why learn R?

As stated on the R project website, R is a programming language and environment for statistical computing and graphics. It is highly versatile, extendable, and community-driven.

R does not involve lots of pointing and clicking, and that's a good thing

The learning curve might be steeper than with other software, but with R, the results of your analysis do not rely on remembering a succession of pointing and clicking, but instead on a series of written commands, and that's a good thing! So, if you want to redo your analysis because you collected more data, you don't have to remember which button you clicked in which order to obtain your results; you just have to run your script again.

Working with scripts makes the steps you used in your analysis clear, and the code you write can be inspected by someone else who can give you feedback and spot mistakes.

Working with scripts forces you to have a deeper understanding of what you are doing, and facilitates your learning and comprehension of the methods you use.

R code is great for reproducibility

Reproducibility means that someone else (including your future self) can obtain the same results from the same dataset when using the same analysis code.

R integrates with other tools to generate manuscripts or reports from your code. If you collect more data, or fix a mistake in your dataset, the figures and the statistical tests in your manuscript or report are updated automatically.

An increasing number of journals and funding agencies expect analyses to be reproducible, so knowing R will give you an edge with these requirements.

R is interdisciplinary and extensible

With 10000+ packages[1] that can be installed to extend its capabilities, R provides a framework that allows you to combine statistical approaches from many scientific disciplines to best suit the analytical framework you need to analyse your data. For instance, R has packages for image analysis, GIS, time series, population genetics, and a lot more.

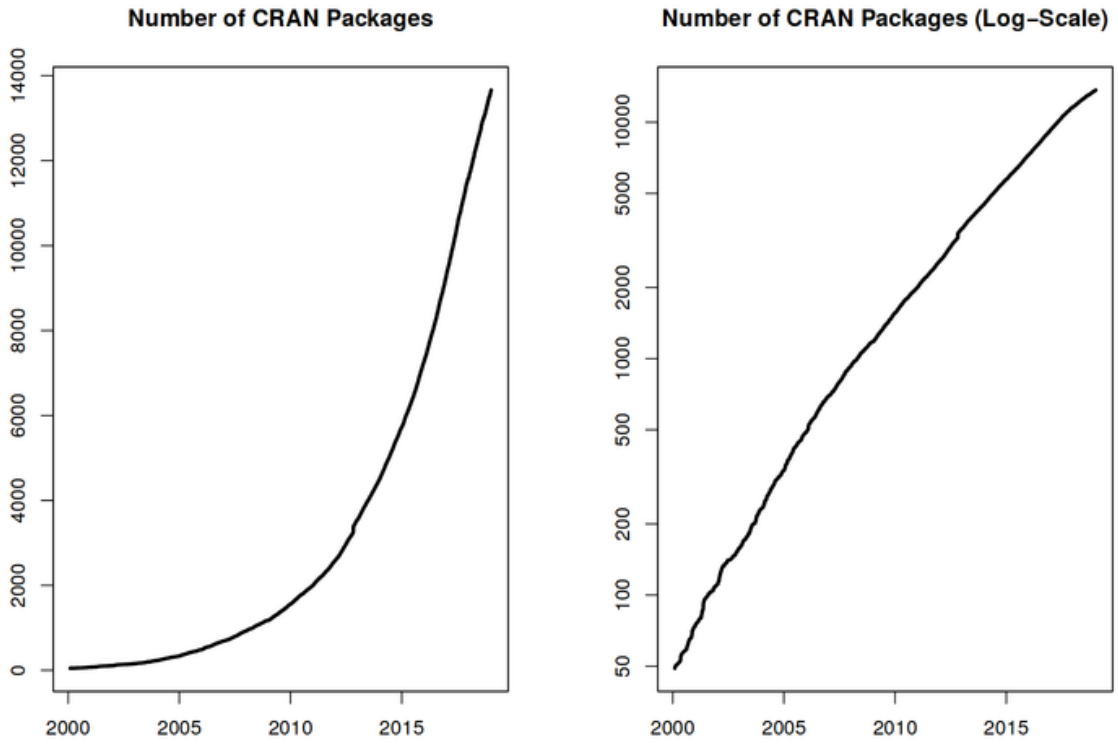

Figure 1: Exponential increase of the number of packages available on CRAN, the Comprehensive R Archive Network. From the R Journal, Volume 10/2, December 2018.

R works on data of all shapes and sizes

The skills you learn with R scale easily with the size of your dataset. Whether your dataset has hundreds or millions of lines, it won't make much difference to you.

R is designed for data analysis. It comes with special data structures and data types that make handling of missing data and statistical factors convenient.

R can connect to spreadsheets, databases, and many other data formats, on your computer or on the web.

[1] i.e. add-ons that confer R with new functionality, such as bioinformatics data analysis.

R produces high-quality graphics

The plotting functionalities in R are extensive, and allow you to adjust any aspect of your graph to convey most effectively the message from your data.

R has a large and welcoming community

Thousands of people use R daily. Many of them are willing to help you through mailing lists and websites such as Stack Overflow, or on the RStudio community. These broad user communities extend to specialised areas such as bioinformatics. One such subset of the R community is Bioconductor, a scientific project for analysis and comprehension "of data from current and emerging biological assays." Another example is R-Ladies, a worldwide organization whose mission is to promote gender diversity in the R community. It is one of the largest organizations of R users and likely has a chapter near you!

Not only is R free, but it is also open-source and cross-platform

Anyone can inspect the source code to see how R works. Because of this transparency, there is less chance for mistakes, and if you (or someone else) find some, you can report and fix bugs.

Sources and References

Most of the materials in this workbook have been adapted from The Epidemiologist R Handbook with some changes made and materials incorporated from other sources. These additional sources are attributed in the chapters they are a part of. The Epidemiologist R Handbook is licensed by Applied Epi Incorporated under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).