

“Introduction to R and the Tidyverse”
PhD Program in Sociology and Social Research
Università di Trento

Mario Quaranta
Department of Sociology and Social Research
Università di Trento
Office: 10 – third floor
e-mail: mario.quaranta@unitn.it

Last update: November 2, 2021

Sessions. The course consists of 4 sessions lasting 3 hours each, held on 9–16–22–29/11/2021, Lab 4.

Office hours. On appointment.

Prerequisites. PhD students are expected to have computing skills and knowledge of other statistical packages. This is not necessary but it helps. Familiarity with basic statistical concepts is assumed.

Description. Among the numerous statistical packages, R stands out for some characteristics that make it a disliked or a popular choice among academics (and not only). R has a bad reputation: it has a steep learning curve, it is designed differently from other statistical packages, help files are often hard to decode, there is a myriad of functions, the same thing can be done in different ways, data management is cumbersome, there are different data structures, issues with memory management, and so on and so forth. Some of these “problems” are eased by the availability of thousands of add-on libraries, which also allow users to carry out advanced analyses. On the other hand, R has characteristics that make it a valuable choice and it should be given a chance (see [here](#)). It is a free and open-source environment that can be installed on any platform. It is very popular (see [here](#)), it is maintained, supported and developed by a vast community which also contributes to blog posts illustrating and solving often-encountered issues (see [here](#) or [here](#)). R is a powerful and flexible tool which allows performing basic algebra to complex simulations requiring computer clusters.

This course will introduce PhD students to the capabilities of R starting from the very basics with the goal of learning the “logic” behind it. This implies that only relatively few things will be taught, while most knowledge will be discovered by the users themselves through applications. The course will be organized in four sessions as follows:

1. Introduction of the core concepts of the R environment and language (as objects, data structures, basic functions, data import and manipulation, basic programming, etc.).

2. Introduction to a collection of libraries – the [Tidyverse](#) – which eases some tasks normally difficult (or annoying) to perform with R base functions. We will discuss the pros and cons of it, trying to focus on what it can be useful for to the beginner’s scope.
3. Introduction to the R graphics capabilities with a focus on how to put in pictures numbers and results.
4. Introduction to how to perform statistical analyses – bivariate and multivariate (or other which are of most interest to participants).

Students at the end of the course will be able to perform statistical analyses using R, but also and most importantly have sufficient knowledge to explore further the software capabilities on their own. It should be born in mind that like any other language (or musical instrument, sport or, in general, skill), R must be practiced to improve and master it. Therefore, during the sessions students will be asked to put in practice what shown with provided examples and data.

Learning strategies. The sessions consist of two parts. In the first part, students will be presented to the core concepts of the session through the illustration of provided examples (i.e. code). In the second part, students will try to apply what learned in the first part with data of their choice. Therefore, tasks will be assigned which have to be completed in-class and/or at home and submitted to the instructor before the following session. The instructor will assist the students and address issues that come up during the in-class practice.

Assessment. Students will be assessed on their participation in the sessions and submission of assignments.

Technical requirements. While it is possible to install R on Android or iOS tablets (see [here](#) and [here](#)) or use online tools (see [here](#)), it is highly recommended that students bring their own laptops, so to avoid possible issues. We will install the software during the first session.

Policies. Attendance of all sessions is mandatory. Absence should be communicated in advance and justified. The use of laptops is strictly limited to academic purposes.

Readings. No readings are necessary to be prepared for each session. There are hundreds of books on R (see [here](#)), useful websites (see one [here](#)) or other freely available documentation (see [here](#)). Below is a list of books/websites that could be used as reference:

- ▷ Dalgaard, P. (2008). *Introductory Statistics with R*. New York: Springer–Verlag.
- ▷ Grolemund, G. & Wickham, H. (2016). *R for Data Science*. Sebastopol: O’Reilly.
- ▷ Healy, K. (2019). *Data Visualization*. Princeton: Princeton University Press.
- ▷ Murrell P. (2011). *R Graphics* [Second edition]. Boca Raton: Chapman & Hall/CRC.
- ▷ *RStudio Cheat Sheets*.
- ▷ Teetor, P. (2011). *R Cookbook*. Sebastopol: O’Reilly.
- ▷ Venables, W. N., Smithand, D. M. & the R Core Team (2020). *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics*. Vienna: R Foundation.