Clara Heinrich
John-F.-Kennedy Institute
Freie Universität Berlin
Email: clara.heinrich@fu-berlin.de

32602 – Introduction to Social Sciences Methods: Statistics, text mining and web-scraping in R

Summer Term 2023
Tuesdays, 12:00-14:00. PC pool, JFKI (Lansstraße 7-9)

Course Description

This course gives a basic introduction to R and Rstudio. We will begin with learning basic data wrangling, manipulation and management techniques. The course then moves on to introduce students to basic statistical concepts, both descriptive and inferential, and their operationalization in the R environment. The second half of the course will focus on how to work with ”text as data”, how to apply various text mining techniques and how to gather large amounts of data through web scraping techniques. The course will generally make use of and acquaint students with a broad range of social science data sources.

Course Objectives

As an introductory course, the seminar will provide students with a general understanding of:
- The R environment (Set-up, functions, packages)
- Basic statistics and their implementation in R

→ Descriptive statistics: Summary statistics, tables, plots
→ Inferential statistics: T-test, correlation, regression analysis and modeling assumptions
- Automated web-data collection
- Finding and accessing various social sciences data sources
- Text analysis

Course Prerequisites

Previous knowledge in statistics is not necessary. Students should bring a general interest in working with data for social sciences research and should be willing to learn, practice and apply statistics.

Recommended Preliminary Readings

Zoë Field, Jeremy Miles, and Andy Field. *Discovering statistics using R.* London: Sage, 2012

→ for R and statistics, uploaded on BlackBoard

Timothy C Urdan. *Statistics in plain English.* New York: Routledge, 2017

→ for statistics, uploaded on BlackBoard

Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis. *Automated data collection with R: A practical guide to web scraping and text mining.* West Sussex: John Wiley & Sons, 2014

→ for web-data collection, uploaded on BlackBoard

Politikwissenschaftliche Statistik mit R. By Christoph Garwe, Philipp Meyer, Laura Brune and Christoph Hönnige, open access [in German]

Follow RStats Question A Day on Twitter for daily tips and tricks in R

Office Hours

You are invited to come and talk to me whenever you feel like doing so and I am happy to discuss anything from questions on homework, the course content, reading or your term paper. Regular office hours are held Tuesdays 4-5pm at my office (room 328, JFKI). Please send me an email beforehand with a preferred time, otherwise you may have to wait. To accommodate students with caring responsibilities or other commitments, I try my best to find flexible solutions for those who cannot come to the regular office hours. In that regard, please send me an email to discuss possible options.

Course Structure & Modes of Teaching

Communication

This course is meant to be a space in which we all learn with and from each other. My aim is to create an atmosphere in which everyone feels comfortable to speak, ask questions and comment on others in a respectful manner. Therefore, I rely on your feedback if you feel like the material is too difficult/ inaccessible for you, if you lack basic knowledge of certain terms or topics, but also if there is something else that makes you feel uncomfortable. Moreover, I would appreciate it if you could let me know if you are unable to attend a session, as this makes it easier for me to plan the sessions in order to make them as lively and interesting as possible.

Recording the seminar sessions is strictly prohibited.

Course Materials

The basic materials of the course are RScripts, which you can find on BlackBoard (Folder: Scripts), and data files (Folder: Data). In addition to the scripts, there will be data homework, which you should complete by using the respective script (Folder: Data Homework) and upload it under assignments on BlackBoard. You also find the syllabus and, if applicable, the readings and slides for the individual sessions on BlackBoard. Scripts and data for each session will be uploaded by 11 am on the day of the session and should be downloaded before/at the beginning of class.

Participation and Examination

For participation, students are expected to actively participate in class. While I do not want to force anyone to speak out loud in class, my aim is that everyone participates according her*his

possibilities. The weekly data homework should be uploaded (as PDF/RScript) **at the latest on Mondays by 11am** on Blackboard assignments.
For full credit, students have to prepare a short student paper pitch in class (10 min maximal) and hand in the finalized student paper (MA: 20 pages, 8000 words main text, BA: 12 pages, 5000 words main text) as a PDF on Blackboard at the latest on **30th September 23:59**. The student paper should develop a research question and answer it using data and techniques learned in class. Discussing your ideas with me during an office hour (prior to the pitch in class) is mandatory.

Class Schedule

Introduction

Session 1: 18.04.2023, Introductory Session

Topics covered:

- What is this course about?
- Why R and how does the program work?
- Where to find what - console, data, packages, scripts, plots, etc.
- Vectors, lists, data-frames
- Which R elements should I know and how do I work with them?
- Loading data, file management with *readr* & co
- Data types and data inspection

Block A - Working with data in the R environment

Session 2: 25.04.2023, Working with data-frames

Topics covered:

- Filtering, sub-setting, inspecting data-frames
- Importing and exporting data-frames

Mandatory readings:

Tidyverse Skills for Data Science, Chapter 1, Introduction to the Tidyverse
Why I don't use the Tidyverse, blog post

Session 3: 02.05.2023, Data manipulation with *dplyr* and baseR

Topics covered:

- Re-coding/ creating variables based on conditions
- "For-loops" and the apply-family

Mandatory readings:

Tidyverse Skills for Data Science, Chapter 3.4, Data Wrangling
RPubs: How to (and how not to) loop in R

Session 4: 09.05.2023, Regular expressions and *stringr*

Topics covered:

- What are regular expressions?
- How to work with regular expressions in R?

Mandatory readings:

Stringr (Your reg-ex bible!)
Tidyverse Skills for Data Science, Chapter 3.7, Working with Strings

---

Session 5: 16.05.2023, Repetition

In this week, we repeat what was covered in the course so far.

Block B - Basic introduction to statistics

Session 6: 23.05.2023, Descriptive statistics

Topics covered:

- Inferential vs descriptive statistics
- Summary measures: Mean, median, mode, variance, standard deviation, etc.
- Distributions

Mandatory readings:

"Why is my evil lecturer forcing me to learn statistics?" Chapter 1 in *Discovering statistics using R*, by Zoë Field, Jeremy Miles and Andy Field, 2012
ModernDive Statistical Background
Chapters 1 and 2 in *Statistics in plain English*, by Timothy C. Urdan, 2017

Further readings:

Tidyverse Skills for Data Science, Chapter 5.5: Descriptive and Exploratory Analysis
Tutorial 7: Descriptive statistics

Session 7: 30.05.2023, Visualization with *ggplot2*

Topics covered:

- Plot types and plot customization

Mandatory readings:

Tidyverse Skills for Data Science, Chapters 4.6 & 4.7, ggplot2: Basics & Customization
"Exploring data with graphs" Chapter 4 in *Discovering statistics using R,* by Zoë Field, Jeremy Miles and Andy Field, 2012

Session 8: 06.06.2023, Inferential statistics

Topics covered:

- Populations and samples
- Causality vs correlation

Mandatory readings:

"Comparing two means", Chapter 9 in *Discovering statistics using R*, by Zoë Field, Jeremy Miles and Andy Field, 2012
Chapters 7 and 8 in *Statistics in plain English*, by Timothy C. Urdan, 2017

Session 9: 13.06.2023, Statistical modelling in R

Topics covered:

- Types of regression analysis
- Model assumptions

Mandatory readings:

"Regression" & "Logistical Regression", Chapters 7 and 8 in *Discovering statistics using R*, by Zoë Field, Jeremy Miles and Andy Field, 2012

Chapter 13 in *Statistics in plain English*, by Timothy C. Urdan, 2017

20.06.2023, No session!

Larger homework combining web data collection, data wrangling and statistics

More details will follow soon

Block C - Web-data collection

Session 10: 27.06.2023, A primer on web-scraping with *rvest*

Topics covered:

- Static vs dynamic web-pages
- Downloading, parsing, extracting from HTML files
- Strategies for trouble shooting

Take a look at *Automated data collection with R: A practical guide to web scraping and text mining* (2014) by Simon Munzert et al.

Block D - Text analysis

Session 11: 04.07.2023, Text as data with *quanteda*

Topics covered:

- Data preparation
- Frequency analysis
- Wordclouds
- Key-word in context analysis

Mandatory reading:

Klaus Krippendorff. *Content analysis: An introduction to its methodology.* Thousand Oaks: Sage publications, 2019

Quanteda tutorials, Sections 1,2,3

Further readings:

Cornel Ban. Content analysis in international political economy. In *The Oxford Handbook of International Political Economy.* Oxford University Press, 2021

Session 12: 11.07.2023, Text as data II

Topics covered:

- Dictionary-based analysis
- Sentiment analysis

Mandatory reading:

Quanteda tutorials, Sections 4,5,6,7

Session 13: 18.07.2023, Student pitches
