Epidemiological Data Analysis using R

Janne Pitkäniemi
Faculty of Social Sciences, Tampere university
Finnish Cancer Registry

janne.pitkaniemi@cancer.fi

Introduction

- Basic properties of R
- Script files
- Data structures and objects
- Data input and output
- Tabulations in R

What is R

Statistical language

- Vast community
- 23 000+ free packages
- Cross plattform compatible (Windows, OS,...)
- can be connected to other languages
- Free and open source
- learning is easy
- excellent graphics

For further information and download: http://www.r-project.org/

More about R

- **Carstensen**, **Bendix**, author. **Carstensen**, B. (2021). **Epidemiology** with R (First edition). Oxford University Press. https://doi.org/10.1093/oso/9780198841326.001.0001
  Carstensen, **Bendix**. 2021. **Epidemiology** with R. First edition.
- Dalgaard, P. *Introductory Statistics with R, 2nd Ed.* Springer, New York, 2008.
- *Statistical Practice in Epidemiology Using R*. An international course, IARC, Tarto, 2020. |http://bendixcarstensen.com/SPE/%7C
- R blog
- Masses of books, articles, websites, etc...

What does R offer for epidemiologists?

- Dynamic publication ecosystem
- Descriptive tools
  - Versatile tabulation
  - High-quality graphics
- Analytic methods
  - Basic epidemiologic statistics
  - Generalized linear models and their extensions
  - Event History/Survival analysis methods
  - Other ...

These are provided by SAS and Stata, too, so why R?

Many features of R are more appealing in the long run.

R ecosystem

- The R language (R)
- R package system (CRAN)
- Development tool/environment (RStudio)
- Dynamic documentation/publishing (quarto)

R language - basis

CRAN - R community

- Packages are collections of new commands, a.k.a. functions, that are developed and shared by the worldwide R userbase.
- base R containes 15 packages
- While CRAN is the most popular package archive for R tools.
- Functions and packages are developed in response to identified needs
- If your needs are unmet by base R, there's likely a package for it
- Altogether, there are more than 20,000 packages on CRAN, alone
- In epidemiology most useful packages are survival and Epi.
- There are tens of thousands of unpublished packages
- Entire ecosystems of packages exist (*Tidyverse)*

Rstudio - script development

Quarto - dynamic publishing

- **Quarto** is an open-source system designed for creating interactive and reproducible documents. Here are some key points about Quarto:
- **Markdown-Based**: Quarto documents are authored using Markdown, making it easy to format text and include code and visualizations.
- **Versatile Output**: You can generate documents in various formats, including HTML, PDF, and presentations.
- **Integration**: Quarto supports integration with RStudio, and other editors, allowing for a seamless authoring experience.
- **Dynamic Reports**: You can create dynamic reports and interactive dashboards using Quarto, making it suitable for scientific and technical writing.

```
[1] 2
```