Introduction to biostatistical computing

arthur.allignol@uni-ulm.de

```r
a <- 1
```

Reproducible Research: A General Definition

Research results are *replicable* if there is enough information to enable an independent researcher to make the same findings using the same procedures.

- This is the aim of science
- Concept closely related to the concepts of replicability and generalisation

This definition is too general for us. We will aim for *reproducible analysis*, i.e.,

“the data and code used to make a finding are available and are sufficient for an independent researcher to recreate the findings”

Reproducible Analysis

What role can we play as (bio)statistician in reproducible research?

Reproducible Analysis

What role can we play as (bio)statistician in reproducible research?

We can at least aim for *reproducible analysis*

- The data and code used to make the findings are available and are sufficient for an independent researcher to recreate the findings
- I.e., from the raw data to the publications in well documented steps

That ain't full reproducibility, but it's already something

Introduction

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.
— D. Donoho

- The concept of reproducible research is based on the idea of *literate programming* such that the logic of the analysis is clearly represented in the final product by combining computer code/programs with ordinary human language
  $\Rightarrow$ Combine analysis code and report

Data analysis work

1. Data cleaning
    - Data entry errors
    - Missing data
    - Recoding
2. Data transformation
    - Transform variables
    - Create new variables
    - Reshape the data entirely to fit some models
3. Statistical analysis (incl. tables, graphs)
4. Statistical report; publication

Research Work, e.g., Master thesis

1. A new model (some maths)
2. Program new model (in R)
3. Simulation study to see whether asymptotic results translate in the real world (small sample properties)
4. Illustrate the usefulness of the new approach with some data
5. Publication; talks

Reproducible Analysis: Barriers/Disadvantages

- Awareness
- Find the right tools
- Learn new tools
- More work from the start

Reproducible Analysis: Barriers/Disadvantages

- Awareness
- Find the right tools
- Learn new tools
- More work from the start

Benefits outweigh the disadvantages

Reproducible Analysis: Benefits

For science:

- Reproducibility is a key part of scientific inquiry
- Reproducibility permits to evaluate scientific claims
- Avoids effort duplication and encourage cumulative knowledge development

Examples:

- A colleague wants to try out your new model
- Somebody wants to compare your approach to hers in your simulation setting

Reproducible Analysis: Benefits

For you:

Better work habits

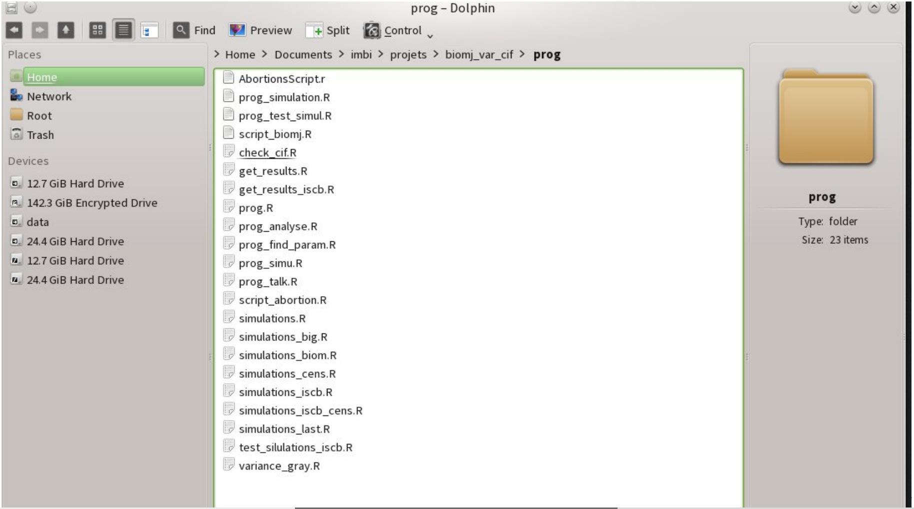
- Making a project reproducible from the start encourages you to use better work habits
  - better organisation
  - better code quality if you think that somebody might actually have a look at it

Examples:

- Perform some supplementary analyses within the review process of a paper

Reproducible Analysis: Benefits

For you:



AbortionsScript.r
prog_simulation.R
prog_test_simul.R
script_biomj.R
check_cif.R
get_results.R
get_results_iscb.R
prog.R
prog_analyse.R
prog_find_param.R
prog_simu.R
prog_talk.R
script_abortion.R
simulations.R
simulations_big.R
simulations_biom.R
simulations_cens.R
simulations_iscb.R
simulations_iscb_cens.R
simulations_last.R
test_silulations_iscb.R
variance_gray.R

Type: folder
Size: 23 items

Reproducible Analysis: Benefits

For you:

Changes are **a lot** easier

- Data analysis is done but the doctors messed up with Z. They are sending you a new data set
- Z is included in a lot of regression models
- You did the data transformation in excel $\rightarrow$ no trace of it
- You copy/pasted the results of the regression models from the R/SAS console in a Word document
- Worst-case scenario: you used point and click software (e.g., SPSS)
- It’s Friday 5pm, the results **have** to be available on Monday 8am

$\Rightarrow$ Have a nice weekend!

Reproducible Analysis: Benefits

Better teamwork

- If an independent researcher can reproduce your analysis, so can a collaborator
- That applies to both current and future collaborators

Higher research impact

- Reproducible research more useful for other researchers

Reproducible Analysis: Benefits

Better teamwork

- If an independent researcher can reproduce your analysis, so can a collaborator
- That applies to both current and future collaborators

Higher research impact

- Reproducible research more useful for other researchers

Some more pragmatic reasons

- We use computers more and more; Quality Assurance processes from the software engineering will come to us
- **My** prediction is that reproducible analysis will become mandatory in the close future

The Tools of Reproducible Research

- A good statistical software that actually permits to write code: **R**, **SAS**, Stata
- *Literate programming*: Combine code with text in a single file
  - ODS (Output Delivery System) in SAS
  - **Sweave** and **knitr** packages in R
- Markup languages, e.g., LaTeX, HTML, Markdown
- A good editor (e.g., Rstudio)
- Some good programming practice, i.e., comments, meaningful variable names

The Tools of Reproducible Research

- A good statistical software that actually permits to write code: **R**, **SAS**, Stata
- *Literate programming*: Combine code with text in a single file
    - ODS (Output Delivery System) in SAS
    - **Sweave** and **knitr** packages in R
- Markup languages, e.g., LaTeX, HTML, Markdown
- A good editor (e.g., Rstudio)
- Some good programming practice, i.e., comments, meaningful variable names

Optional but very useful

- Version control software
- Shell scripting abilities (for bigger projects)

Statistical Software

The way you interact with R, SAS, Stata or other programming languages has benefits for reproducible research

- Interaction with the language by explicitely writing down your steps as source code
- A lot better than point and click programs for reproducible research
    - Your steps are usually lost when you click around to fit models
- *Literate programming* only possible with source code

Literate Programming

“Literate programming is an approach to programming introduced by Donald Knuth (1970s) in which a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated”, i.e., *writing documentation containing computer code*

Literate Programming

“Literate programming is an approach to programming introduced by Donald Knuth (1970s) in which a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated”, i.e., *writing documentation containing computer code*

- For statistician it means being able to
  - combine programming code with report text (article, presentation) in a single self-documenting file
  - document the code and its results, including interpretation of the results
  - allow an analysis to be rerun and the report (article, presentation) to be re-typeset by running a single command

Literate Programming

In R

- The **Sweave** package: Embed R code in LaTeXfiles. A pass through Sweave converts the file to a `.tex` file, in which code, tables, graphics are included
- The **knitr** package: An evolution of **Sweave**. **knitr** adds some useful features and permits additionally to embed R code in html and markdown documents
- Plenty of useful packages that creates markup tables from, e.g., model fit

In SAS

- SASweave: SAS code in LaTeX
- ODS: Output SAS “stuffs” in various file format (pdf, rtf, html, ...)

Good coding practice

Comments

- Text in the code that are not compiled/executed by the program
- Explain what difficult pieces of code do, e.g.,

```
for (i in seq_along(times)) {
    dna[cbind(ii, ii, i)] <- -(.rowSums(nev[, , i],
        nrisk[i, ]
}
```

Meaningful variable names

- `temp1`, `temp2`, etc. are not good
- `Horrible_Disease` with value `Yes` and `No` is better

Header

- Write as a comment the aim of the program, your name and email at the beginning of the file

A good Editor

- Auto-completion is an extremely useful feature, especially when using meaningful (long) variable names.
- Rstudio is a good R editor
  - Good interaction with markup languages
  - Interaction with version control systems
  - Using **Sweave** or **knitr** is made easy

Advanced Tools: Version Control and Shell

Version Control

- Practice that tracks and provides control over changes to files (e.g., source code)
- Example: Git

Shell scripts

- When projects get big, having everything in one file is actually a problem
- Shell scripts permit to “glue” several analyses together
- Example: A single shell script executes separated R programs

But...

Perfect reproducibility is not achievable

- Differences in
  - Operating systems
  - Processors
  - Software and package versions
  - . . .

Simulation Studies

What is a simulation study?

Simulation A numerical technique for conducting experiments on the computer

Monte Carlo simulation Computer experiment that involves random sampling from probability distributions

- **Extremely** useful in statistics
- When a statistician talks about simulation, it usually means Monte Carlo simulation

Simulation Studies

Why do we do simulation studies?

- Properties of statistical methods should be established,
- but analytical derivations rarely possible
- Large sample approximations of these properties are possible
- but at the end of the day, we need to evaluate the methods in (finite) sample sizes that are likely to be seen in practice
- Analytical results usually require assumptions
- We also need to know what happens when these assumptions are violated
    - Extremely difficult to do analytically

Simulation Studies

Simulation studies permit, under various conditions, to answer questions like

- What is the bias of an estimator in finite samples
- Is this estimator still consistent under departures from the assumptions
- How does a new statistic compare to competing ones in terms of bias, precision
- Does the procedure for constructing a confidence interval for a parameter achieve the nominal level of coverage

This questions cannot usually be answered analytically

Simulation Studies

How does that work?

1. Generate S independent data sets under the condition of interest
2. Compute the numerical value of the statistic of interest T for each data set. We obtain $\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_S$
3. If S large enough (say, 1000), summary statistics across $\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_S$ is a good approximation to the true sampling properties of the statistic under the conditions of interest

Simulation Studies

What we will learn?

- Generate realistic data sets
- Summarise simulation results in some intelligible way
- If time permits: Parallelisation