Introduction to biostatistical computing

arthur.allignol@uni-ulm.de

Table of Contents

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

Reproducible Analysis

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

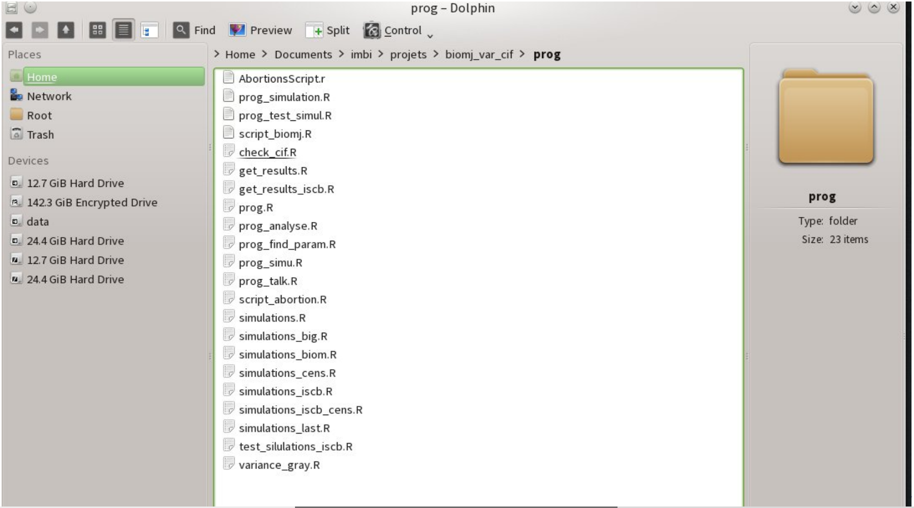
- Making a project reproducible from the start encourages you to use better work habits
  - better organisation
  - better code quality if you think that somebody might actually have a look at it

Examples:

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

Reproducible Analysis: Benefits

For you:



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 LaTeX files. 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], dim(nev)[1],
                                        dim(nev)[1], FALSE)) /
        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

Good file organisation

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
  - ...

Table of Contents

Real Life Data

```r
library(fortunes)
fortune("Tolstoy")

##
## Happy families are all alike; every unhappy family is
## unhappy in its own way.
## Leo Tolstoy
##
## and every messy data is messy in its own way - it's easy to
## define the characteristics of a clean dataset (rows are
## observations, columns are variables, columns contain values
## of consistent types). If you start to look at real life
## data you'll see every way you can imagine data being messy
## (and many that you can't)!
##    -- Hadley Wickham (answering 'in what way messy data
##       sets are messy')
##       R-help (January 2008)
```

	A	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE
1		Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert	Mittelwert
2		PTR50	PTR50	PTR50	PTR50	BasePost	BasePost	BasePost	BasePost	BasePost	BasePost
3		ACMmean	ABPmean	Puls	CO2	ACEsyst	ACEdiast	ACEmean	ACE_RI	ACE_PI	ACMmean
4		cm/s	mmHg	Schläge/min		cm/s	cm/s	cm/s			cm/s
5											
6											
7											
8											
24	M16	76.371167	82.266667	80.72	38.86	21.662667	7.9630833	12.978992	0.6304667	1.0567333	74.523167
25	M17	63.063	88.45	61.313333	44.806667	22.201667	6.6906583	11.139333	0.6969	1.3963333	54.4775
26	M18	58.430167	80.633333	59.263333	36.986667	23.46575	5.98675	11.139975	0.7432	1.5663333	59.007667
27	M19	60.676	60.356667	85.326667	33.746667	22.81125	5.2969583	10.882025	0.7523	1.5543333	58.070833
28	M20	ND	ND	ND	ND	0	0	0	0	0	ND
29	M21	ND	ND	ND	ND	0	0	0	0	0	ND
30	M22	ND	ND	ND	ND	0	0	0	0	0	ND
31	M23	ND	ND	ND	ND	0	0	0	0	0	ND
32	C01	63.2555	68.05	62.553333	40.116667	24.255	7.417025	11.974783	0.6934333	1.409	63.1785
33	C02	78.398833	75.466667	64.183333	30.736667	25.461333	9.5011583	14.1372	0.6257667	1.1303333	83.108667
34	C03	81.851	65.386667	66.006667	40.083333	23.69675	8.90505	13.329983	0.6204667	1.12	81.260667
35	C04	44.826833	60.243333	71.69	40.226667	23.69675	8.8088	12.192308	0.6270333	1.2216667	42.106167
36	C05	60.881333	69.37	57.713333	40.006667	25.050667	5.79425	10.963517	0.7676667	1.76	61.664167
37	C06	61.895167	64.723333	56.256667	44.95	24.511667	2.8602292	9.50565	0.8815667	2.337	60.188333
38	C07	67.272333	82.496667	68.973333	31.786667	20.46275	7.546	11.96965	0.6287667	1.0839667	63.961333
39	C08	79.8875	61.253333	63.253333	39.96	22.11825	6.7554667	11.578875	0.6934333	1.3286667	86.329833
40	C09	56.081667	69.716667	72.756667	35.846667	20.798342	5.6672	11.413967	0.7303333	1.3773333	57.904
41	C10	70.968333	70.63	59.98	33.92	24.210083	8.3211333	12.3816	0.6562333	1.3261	69.6465
42	C11	65.4885	57.193333	54.646667	39.116667	22.913917	6.966575	11.174625	0.6917333	1.4503333	63.550667
43	C12	71.212167	68.53	57.23	33.403333	24.172225	4.4589417	9.51335	0.8033333	2.051	68.748167
44	C13	62.498333	67.456667	68.316667	35.743333	21.5985	5.026175	8.9525333	0.7446667	1.78	59.931667
45	C14	55.388667	96.22	59.04	34.883333	18.501817	6.5822167	10.1409	0.6438667	1.1763333	55.183333
46	C15	64.397667	69.9	58.443333	39.443333	18.057142	6.2703667	10.144108	0.6583667	1.1783333	59.700667
47	C16	50.832833	73.603333	53.75	37.706667	23.31175	9.4556	13.023267	0.5929333	1.0642	48.715333
48	C17	56.877333	64.573333	62.143333	44.436667	21.278308	7.2688	11.825275	0.6631667	1.1966667	56.646333
49	C18	71.853833	99.513333	71.756667	33.19	23.472167	9.3413833	14.218692	0.6012667	0.9945667	69.081833
50	C19	56.120167	79.136667	79.21	35.913333	20.748933	6.1086667	9.9721417	0.7038333	1.4726667	52.206
51	C20	ND	ND	ND	ND	0	0	0	0	0	ND
52											
53											
54	Mittelwert M	63.362895	78.68807	69.080175	38.029123	23.416612	6.6535396	11.807241	0.7116263	1.4604316	59.353491
55	Standardab M	11.028062	11.74354	11.654379	2.8089418	2.2251624	2.1207966	1.7868907	0.0849892	0.3577957	10.566291
56											
57	Mittelwert C	64.209895	71.761228	63.57386	37.445789	22.542966	7.0028967	11.495391	0.6856772	1.3925351	63.321693
58	Standarab C	9.889746	11.079763	7.0194203	3.9793285	2.1215214	1.8141646	1.5395569	0.0741666	0.3570025	11.328758

Real Life Data

The problems

- Bad structure
- Non ascii characters (e.g., Umlaut)
- Variable names with spaces; on several lines, ...
- Colour coding
- No consistent definition of missing values
- Free text
- (Wrong input, e.g., person who die twice)
- ...

Real Life Data

The problems

- Bad structure
- Non ascii characters (e.g., Umlaut)
- Variable names with spaces; on several lines, ...
- Colour coding
- No consistent definition of missing values
- Free text
- (Wrong input, e.g., person who die twice)
- ...

Often data need to be transformed/reshaped for fitting a particular model

Real Life Data

The tools needed

- Read data
- Work with strings/characters variables
- Factors and dates
- Data manipulation, reshaping, merging

Table of Contents

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 $T_1, T_2, \ldots, T_S$
3. If S large enough (say, 1000), summary statistics across $T_1, T_2, \ldots, T_S$ 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

Table of Contents

Further Topics

Graphics

Why do we need graphics?

- Data exploration
  - Look for trends and/or associations between variable
  - The first step before modelling
  - $\rightarrow$ quick and dirty graphs
- Check assumptions of statistical models
- These graphics are usually for you

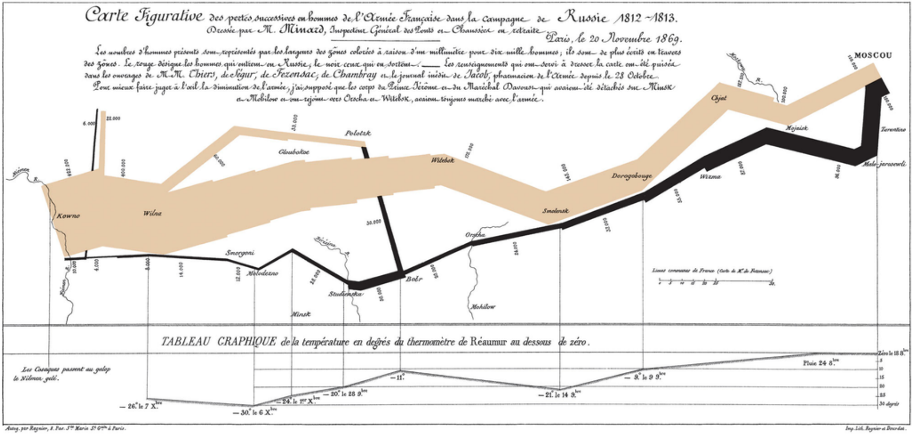
Further Topics

Graphics

Why do we need graphics?

- Information visualisation/Communication
    - Need a lot of polishing
    - Iteration is crucial
    - Think about where you present the graphics, e.g, colour, line thickness for a beamer presentation

Minard's Flow Map



Further Topics

- Debugging
- Something you might want to hear about?