Reproducible data analysis

Daniel Hammarström

2019-10-06

Content

- Defining replication and reproducibility
- Identifying challenges facing the data analyst in the sport sciences
- How to perform a reproducible data analysis - tools and workflow

Replication

Replication is the degree to which scientific findings can be repeated using:

- independent data,
- independent research-groups, laboratories
- different methods, instrumentations. . .

(Peng, Dominici, and Zeger 2006)

The replication crisis

Many scientific results are not possible to confirm. Ioannidis (2005) and others points out that scientific results most often are not true:

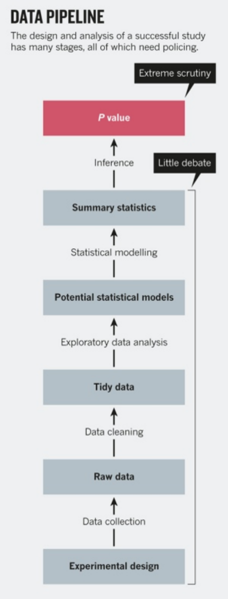
- Smaller studies are more likely to produce untrue conclusions
- Large amount of hypotheses produces large amount of false positives
- Low degree of scientific stringency produces more false claims
- E.g. financial interests increases the risk of false claims
- New fields more likely to produce false conclusions

Reproducibility

- Reproducibility is the degree to which the same conclusions/interpretations can be made using a single dataset.
- A reproducible data analysis can be scrutinized on the level of the data, methods and documentation.
- This can be seen as a “minimum standard” in reporting a study when replication is not feasible (Peng, Dominici, and Zeger 2006).
- Reproducability is a way to bridge the gap between full replication to a non-replicable study (Peng 2011).
- A reproducible study report is a well documented and explicit data-analysis.

Why make data analysis explicit?

Before calculating a p-value (and making inference), most of the "scientific black-box activities" takes place.



DATA PIPELINE

The design and analysis of a successful study has many stages, all of which need policing.

The data analysis pipeline – A crucial part of every scientific project

- Not easy to fully implement. . .
- “From sample to data-point” descibed prior to data collection
- The goal is to make informed *a priori* decisions based on the question not a specific result (avoids bias (Ioannidis 2005))

The data analysis pipeline

- Experimental design
- Data collection
- Raw data
- Tidy data
- Data cleaning
- Exploratory data analysis
- Statistical models
- Summary statistics
- Inference (p-values)

(Leek and Peng 2015)

Data collection

- Errors during data collection could be human and non-human errors
- Use data validation when entering data manually into a spreadsheet
- Avoid “black-box” data-capturing to avoid non-human errors

Raw data

- Raw data is not processed, manipulated or transformed from other data.
- This is however relative . . .
- In order to use raw data in an analysis, many processing steps may be needed

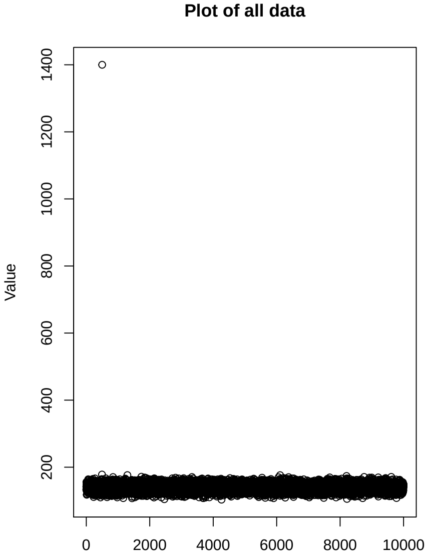
Data cleaning

- Data cleaning can be performed on your *raw data* to make it more suitable for analysis.
- Data cleaning is the process of formatting and sorting data into a suitable table-like format that is readable for you and your software.
- Data cleaning also identifies missing data or incorrectly entered data

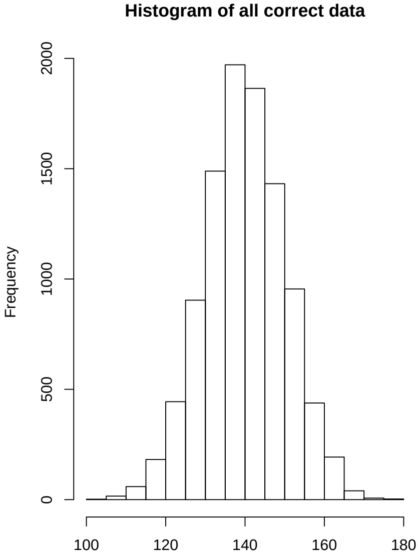
Data cleaning

- To identify e.g. problems in data collection (data entry or recording) we can use summary statistics
- A data set contains 10000 observations. We expect an average value around 140 units, the observed average is 140.06. The standard deviation is 16.16, the data is quite concentrated. However the maximum is 1400, this means we should be careful with making the assumption that the data is correct as this number is about 10 times the expected value.
- Graphical analysis of the data is useful as it summarises many aspects of the data

Graphical method for identifying data problems



Plot of all data



Histogram of all correct data

The goal of data cleaning -> Tidy data

- Tidy data: each row is an **observation** and each column is a **variable** and each cell of the table contains **values**

Variable	Variable	Variable	Variable	Variable
Value	Value	Value	Value	Value
Value	Value	Value	Value	Value
Value	Value	Value	Value	Value
Value	Value	Value	Value	Value

Table 1: Example of tidy data

Tidy data?

Participant	StrengthWeek1	StrengthWeek2	StrengthWeek3
FP1	120	125	140
FP2	130	140	145
FP3	140	145	130

Table 2: Example of tidy data?

Tidy data?

Participant	time	strength
FP1	week1	120
FP1	week2	125
FP1	week3	140
FP2	week1	130
FP2	week2	140
FP2	week3	145
FP3	week1	140

Table 3: Example of tidy data?

Why aim for tidy data?

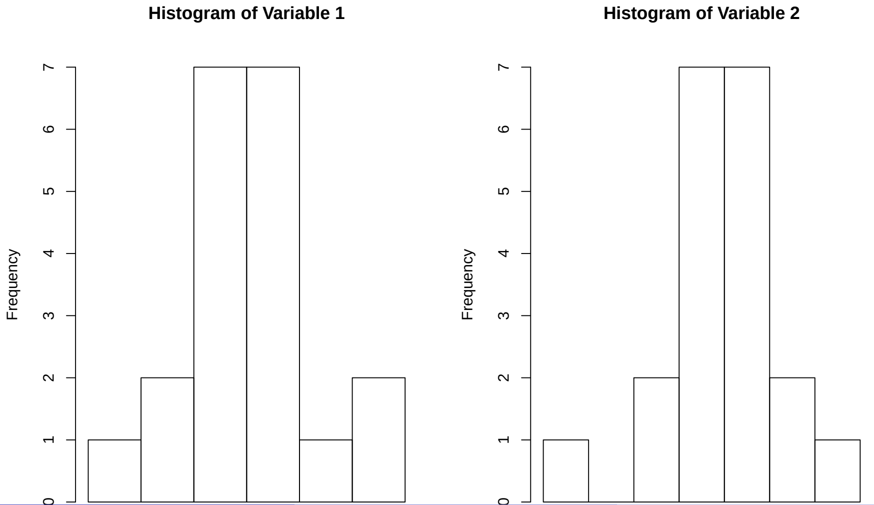
- Tidy data (Wickham 2014) builds upon a standard practice for organizing data
- Makes data exploration and analysis easier
- Standard for easy description and sharing of data
- Tidy data is the standard input in many statistical modeling routines in R, SPSS, SAS and STATA

Exploratory data analysis

- When you have a data analysis plan, the purpose of the exploratory data analysis is to check assumptions about your data.
- Graphical methods is handy!
- Tidy data makes your exploratory analysis easy

Exploratory data analysis

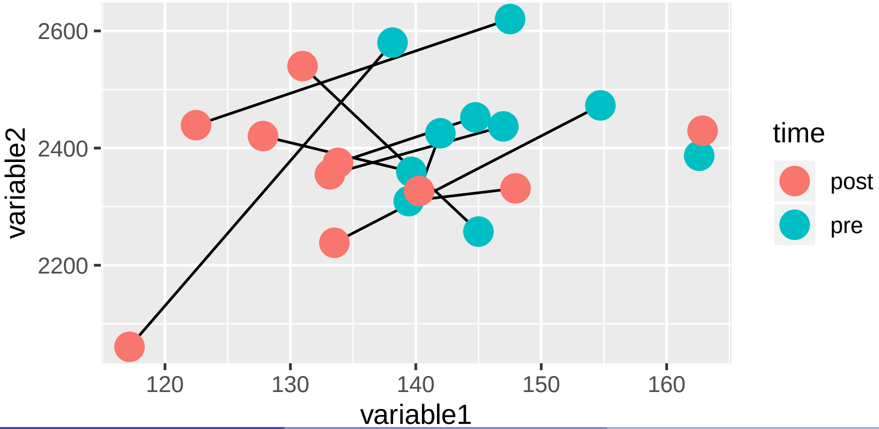
Checking assumptions using graphical methods



Exploratory data analysis

Checking assumptions using graphical methods, more information on one plot!

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```



How to do explicit data analyses – Prior to, or after data collection

Make a data analysis plan based on the goal of the study:

- Describe variables that are going to be collected, this is your **codebook**.
- Create specific data storage for grouped data variables (e.g. spreadsheets).
- Describe how *raw-data* is transformed to *analytic data* (e.g. force-curves from isokinetic testing is converted to maximal torque).
- Describe how summary statistics are to be calculated
- Make *dummy*-tables and figures
- Describe what tests you need to draw inference

Setting up an analysis workflow

Use a standard folder system that contains all data, analyses and reports. The whole project is contained in one *root folder*.

- Write a **codebook.txt** containing all variables in the dataset, how is data stored and accessed.
- **./data/**: contains all your data, e.g. spreadsheets with data from specific tests
- **./analysis/**: contains all analysis-files, e.g. SPSS-files that does analyses
- **./output/**: contains results from analyses, figures etc.

Working with spreadsheets

- Spreadsheets are mainly suitable for data storage and entry.
- Making statistical analyses and or figures in spreadsheets is not adviced
- If you plan to make analyses in spreadsheets, store data in one sheet and make analyses and data manipulation in another spreadsheet.

(Broman and Woo 2017)

Guidelines for working with spreadsheets

- The goal is to make human- and computer-readable spreadsheets.
- Consistency, use the same variable names, coding etc for all data storage (e.g. *Male*, *Female*, *M*, *F* etc.)
- Good naming strategy, do not use spaces (e.g use *strength_week_1* instead of *Strength week 1*), avoid special characters (£, @, €).
- Write dates like YYYY-MM-DD
- Don't leave cells empty, don't comment if you have missing values, use a specific variable for comments
- Keep the data tidy, in a rectangle.
- Do not use color, different fonts etc.
- Use data validation when entering data
- No summary statistics or calculations!
- Make a separate **codebook** (data dictionary)

(Broman and Woo 2017)

The codebook

- Variable names
- Explanation of the variable name, how was it recorded/collected
- Unit of measurement
- Expected values min-max

Why work with reproducible methods

- You may want to collaborate
- Your most frequent collaborator is yourselfe in the future
- Standard practices helps your collaborators (you!) to pick up where you left off

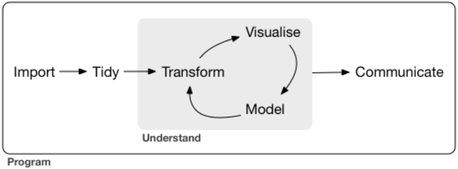
Don’t do the data analysis, tell the computer how to! Learn how to program!

- *Scripting* a data analysis will increase the potential for full reproducibility
- A sucessfull data analyst is **lazy**, tell the computer to do boring stuff!

Don't do the data analysis, tell the computer how to! Learn how to program!

- Investing time in a programming language is investing in a generic skill
- The demand for data-analysis skills will likely increase (especially in sport science)

Data anlysis



What to learn if you want to be a better analyst/researcher

- R is a good place to start if you want do data analysis
- R has a large user community, easy to find help
- R is free, open source and users contribute with programs (this is not true for e.g. SPSS, SAS. . .)
- R is implemented in many graphical user-interface programs like Jamovi
- Using R with RStudio makes a very productive analyst environment
- With R and resources like R Markdown you can write reports with based on your data.
- Many statistical courses, books and resources use R

References

Broman, Karl W., and Kara H. Woo. 2017. "Data Organization in Spreadsheets." Journal Article. *The American Statistician*, 0–0. https://doi.org/10.1080/00031305.2017.1375989.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." Journal Article. *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Leek, J. T., and R. D. Peng. 2015. "Statistics: P Values Are Just the Tip of the Iceberg." Journal Article. *Nature* 520 (7549): 612. https://doi.org/10.1038/520612a.

Peng, R. D. 2011. "Reproducible Research in Computational Science." Journal Article. *Science* 334 (6060): 1226–7. https://doi.org/10.1126/science.1213847.

Peng, R. D., F. Dominici, and S. L. Zeger. 2006. "Reproducible Epidemiologic Research." Journal Article. *Am J Epidemiol* 163 (9): 783–9. https://doi.org/10.1093/aje/kwj093.

Wickham, Hadley. 2014. "Tidy Data." Journal Article. *Journal of Statistical Software; Vol 1, Issue 10 (2014)*. https://www.jstatsoft.org/v059/i10 http://dx.doi.org/10.18637/jss.v059.i10.