

THE EMACS ORG-MODE

Reproducible Research and Beyond

Andreas Leha

Department for Medical Statistics
University Medical Center Göttingen

Outline

Reproducible Research

Existing Tools for Reproducible Research

Org-mode

Summary

Orientation

Reproducible Research

Existing Tools for Reproducible Research

Org-mode

Summary

What is Reproducible Research?

Possible Definition

a piece of reproducible research is an article that provides readers with all the materials that are needed to produce the same results as described in the publication

(Hothorn, Held, and Friede, 2009)

What is Reproducible Research?

Possible Definition

a piece of reproducible research is an article that provides readers with all the materials that are needed to produce the same results as described in the publication

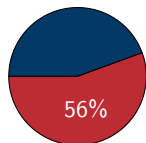
(Hothorn, Held, and Friede, 2009)

A piece of reproducible research is usually not...

methods section + published data

Nature Genetics (2005/06)

(Ioannidis et al., 2009)



not reproducible

What is Reproducible Research?

Possible Definition

a piece of reproducible research is an article that provides readers with all the materials that are needed to produce the same results as described in the publication

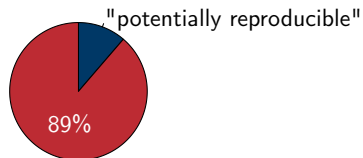
(Hothorn, Held, and Friede, 2009)

A piece of reproducible research should ideally contain...

methods section + published data + code

Biometrical Journal (Vol.50)

(Hothorn, Held, and Friede, 2009)



Bioinformatics (Vol.26)

(Hothorn and Leisch, 2011)

- ▶ Better but similar results

What is Reproducible Research?

Possible Definition

a piece of reproducible research is an article that provides readers with all the materials that are needed to produce the same results as described in the publication

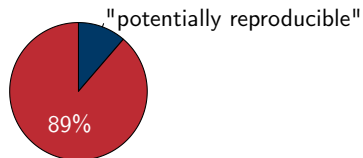
(Hothorn, Held, and Friede, 2009)

A piece of reproducible research should ideally contain...

methods section + published data + code + parameters + ...

Biometrical Journal (Vol.50)

(Hothorn, Held, and Friede, 2009)



Bioinformatics (Vol.26)

(Hothorn and Leisch, 2011)

- ▶ Better but similar results

Why to use Reproducible Research

- ▶ Benefits for the researcher herself

In the mid-1980s, we realized that our laboratory's researchers often had difficulty reproducing their own computations without considerable agony.

(Schwab, Karrenbach, and Claerbout, 2000)

- ▶ Benefits for others

- ▶ Precise 'description' of methods
- ▶ Easy re-use of applied methods
- ▶ No *forensic bioinformatics*

(Baggerly and Coombes, 2009; Ioannidis et al., 2009)

Barriers for Reproducible Research

(Banks, 2011)

- ▶ Deliberate non-reproducibility
 - ▶ Vagueness to cover potential mistakes
- ▶ External reasons
 - ▶ Ownership of the data
 - ▶ Collaboration partners
- ▶ Perception of effort
 - ▶ Discipline
 - ▶ Resources
 - ▶ Change of work flow

Orientation

Reproducible Research

Existing Tools for Reproducible Research

Org-mode

Summary

Existing Tools for Reproducible Research

- ▶ ReDoc (Schwab, Karrenbach, and Claerbout, 2000)
GNU make rules synchronize code and output
- ▶ Sweave (Leisch, 2002)
interwoven *R* and \LaTeX by means of *literate programming*
- ▶ Compendium (Gentleman and Temple Lang, 2007; Gentleman et al., 2005)
scientific paper as *R*-package (including data, code)
based on Sweave
- ▶ Org-mode

Orientation

Reproducible Research

Existing Tools for Reproducible Research

Org-mode

Summary

What is Org-mode?

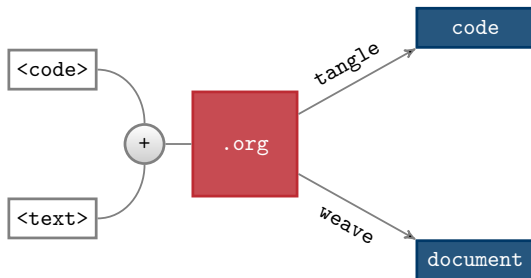
Org-mode is for keeping notes, maintaining ToDo lists, doing project planning, and authoring with a fast and effective plain-text system.

<http://orgmode.org/>

- ▶ Major mode of emacs
- ▶ File format
- ▶ Created in 2003 by Carsten Dominik
- ▶ Current Version 7.7; maintainer Bastien Guerry
- ▶ Very active development

Org-mode and Reproducible Research

- ▶ Through [Org Babel](#), a *literate programming* (Knuth, 1984) system
- ▶ Written by Eric Schulte and Dan Davison



A Simple Example

The Data

We first **generate** some data:

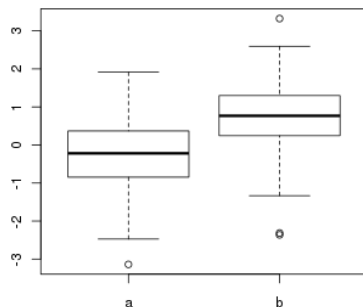
```
a <- rnorm(100, mean=0.0)
b <- rnorm(100, mean=0.8)
```

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Descriptive Values

	mean	sd
a	-0.21	1.04
b	0.71	0.92

Visualization



Analysis

A t-test shows that the means are significantly different (p-value:).

Skip Screenshots

A Simple Example

```
File Edit Options Buffers Tools Org Tbl Help
#+TITLE: A Simple Example
#+BABEL: :session *orgmode* :exports results
#+OPTIONS: toc:nil num:nil

* The Data...

* Descriptive Values...

* Visualization...

* Analysis...

* Using Noweb :noexport:...

* Latex Options :noexport:...

--:-- Demo.org All (1,0) Git:master (Org Fill)-----
```


A Simple Example

```
File Edit Options Buffers Tools Org Tbl Help
#+TITLE: A Simple Example
#+BABEL: :session *orgmode* :exports results
#+OPTIONS: toc:nil num:nil

* The Data

We first *generate* some data:

#+source: generate_data
#+begin_src R :exports code
  a <- rnorm(100, mean=0.0)
  b <- rnorm(100, mean=0.8)
#+end_src

* Descriptive Values...

* Visualization
--:-- Demo.org      Top (5,0)      Git:master (Org Fill)-----
```

A Simple Example

```
File Edit Options Buffers Tools Org Tbl Help
* Descriptive Values

#+begin_src R :colnames yes :rownames yes
  require("plyr")
  calcDV <- each(mean,sd)
  dv <- rbind(a=calcDV(a),b=calcDV(b))
  round(dv,2)
#+end_src

#+LaTeX: \pagebreak

* Visualization...

* Analysis...

* Using Noweb                                     :noexport:...

--:-- Demo.org      8% (17,0)   Git:master (Org Fill)-----
```

A Simple Example

```
File Edit Options Buffers Tools Org Tbl Help
* Descriptive Values

#+begin_src R :colnames yes :rownames yes
  require("plyr")
  calcDV <- each(mean,sd)
  dv <- rbind(a=calcDV(a),b=calcDV(b))
  round(dv,2)
#+end_src

#+results:
|   | mean | sd |
|---+---+---|
| a | -0.14 | 1.12 |
| b |  0.73 | 1.02 |

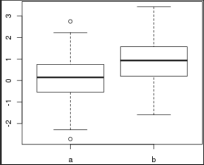
#+LaTeX: \pagebreak

--:**- Demo.org      8% (17,0)   Git:master (Org Fill)-----
```

A Simple Example

```
File Edit Options Buffers Tools Org Tbl Help
* Visualization
#+source: boxplot
#+begin_src R :results graphics :file boxplot.png :width 300 :height 250
  par(mar=c(2,2,0,0)+0.1)
  boxplot(list(a=a,b=b))
#+end_src

#+ATTR_LaTeX: width=7cm
#+results: boxplot
```



```
--:**- Demo.org 17% (35,0) Git:master (Org Fill)-----
```

A Simple Example

```
File Edit Options Buffers Tools Org Tbl Help
* Visualization...
* Analysis
#+source: analyze_data
#+begin_src R :exports none
  p <- t.test(a,b)$p.value
#+end_src

#+results: analyze_data
: 0.000316729293949556

A t-test shows that the means are significantly different
(p-value: src_R[:results raw]{format(p, digits=2)}).

* Using Noweb :noexport:...
* Latex Options :noexport:...
--:**- Demo.org Bot (48,0) Git:master (Org Fill)-----
```

A Simple Example

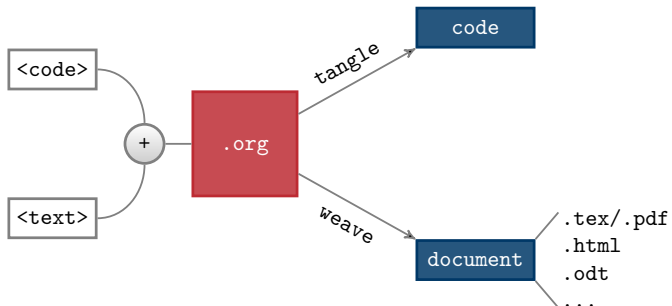
```
File Edit Options Buffers Tools Org Tbl Help
* Visualization...
* Analysis...
* Using Noweb :noexport:
  #+begin_src R :noweb yes
    <<generate_data>>
    <<analyze_data>>
  #+end_src

  #+results:
  : 6.19585741319946e-07

* Latex Options :noexport:...
```

```
--:**- Demo.org Bot (61,0) Git:master (Org Fill)-----
```

Export



A Simple Example

file:///home/andreas/work/reproducible_research/org_mode/pr...

A Simple Example

The Data

We first **generate** some data:

```
a <- rnorm(100, mean=0.0)
b <- rnorm(100, mean=0.8)
```

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Descriptive Values

	mean	sd
a	0.08	0.99
b	0.84	1.08

Visualization

The box plot shows two distributions. The left distribution (a) has a median around 0.08, a mean around 0.08, and a standard deviation of 0.99. The right distribution (b) has a median around 0.84, a mean around 0.84, and a standard deviation of 1.08. Both distributions show outliers above the upper whiskers.

File Bearbeiten Ansicht Einfügen Format Tabelle Extras Fenster Hilfe

Heading 1.title Arial 16,1

A Simple Example

1. The Data

We first **generate** some data:

```
a <- rnorm(100, mean=0.0)
b <- rnorm(100, mean=0.8)
```

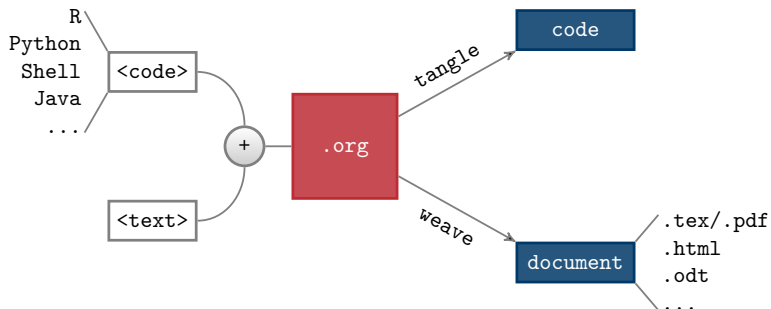
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

2. Descriptive Values

	mean	sd
a	-0.06	1.11
b	0.92	1.04

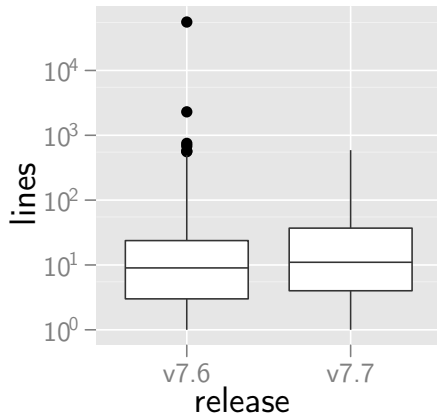
3. Visualization

Various Languages



Various Languages

Example: Commit Sizes in Org-mode v7.6 and v7.7



lrr	mean	sd
v7.6	139.2	2401.96
v7.7	40.48	81.34

A Wilcoxon test shows 'no significant' difference (p-value: 0.082).

Skip Screenshot

Various Languages

Example: Commit Sizes in Org-mode v7.6 and v7.7

```
File Edit Options Buffers Tools Org Tbl Help
** Various Languages                                     :B frame:
*** Example: Commit Sizes in Org-mode v7.6 and v7.7     :B block:
:PROPERTIES:...

#+source: commitstats(FROM="release_7.6", TO="release_7.7")
#+begin_src sh :session none :exports none
  cd ${HOME}/local/emacs/org-mode

  git log --no-merges --oneline \
    --shortstat ${FROM}..${TO} | \
    awk 'NR % 2 == 0'
#+end_src

#+results: commitstats
| 105 files changed | 107 insertions(+) | 107 deletions(-) |
| 1 files changed  | 2 insertions(+)  | 2 deletions(-)  |
| 1 files changed  | 11 insertions(+) | 22 deletions(-) |
| 1 files changed  | 6 insertions(+)  | 6 deletions(-)  |
| 1 files changed  | 1 insertions(+)  | 0 deletions(-)  |
-U:***- 2011_useR.org 36% (1046,0) Git:master (Org Bm Fill)-----
```

More Examples on Org Babel

- ▶ Website on uses of Org Babel
<http://orgmode.org/worg/org-contrib/babel/uses.html>
- ▶ Comparison to Sweave by demo (Eric Schulte)
<http://orgmode.org/worg/org-contrib/babel/uses.html#foo>
- ▶ Tutorial 'Org-mode and R' by Erik Iverson
<https://github.com/erikriverson/org-mode-R-tutorial>
- ▶ Examples reproducible research papers
 - ▶ "Active Document with Org-Mode", Schulte and Davison (2011) on Org-mode itself
<https://github.com/eschulte/CiSE>
 - ▶ "A Model-based Age Estimate for Polynesian Colonization of Hawai'i", Dye (in press) with Setup for Org-mode
<https://github.com/tsdye/hawaii-colonization>

Org-mode for 'beyond'

Some Highlights

- ▶ Note taking
 - ▶ Outlining / Folding
 - ▶ Rearrangement of whole branches
- ▶ ToDo lists / Organizer
 - ▶ Agendas
 - ▶ Scheduling
 - ▶ Mobile apps
- ▶ Tables / Spreadsheet

Orientation

Reproducible Research

Existing Tools for Reproducible Research

Org-mode

Summary

Summary

Using Org-mode for Research Can Give You

- ▶ Reproducibility
- ▶ Plain Text Files
- ▶ Visual User Experience
- ▶ Various Export Formats
- ▶ Various Programming Languages
- ▶ Intuitive Organizer
- ▶ ...

But

- ▶ Reproducibility is still limited by active development
- ▶ Less editing support than Sweave

Disclaimer

Claerbout's principle

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.

(Buckheit and Donoho, 1995)

Disclaimer

Claerbout's principle

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.

(Buckheit and Donoho, 1995)

Credit

This presentation about Org-mode is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software and all the credit goes to the developers.

Literature I

- Baggerly, Keith A. and Kevin R. Coombes (2009). "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology". In: *Annals of Applied Statistics* 3 (4), pp. 1309–1334.
- Banks, David (2011). "Reproducible Research: A Range of Response". In: *Statistics, Politics, and Policy* 2 (1).
- Buckheit, Jonathan B. and David L. Donoho (1995). "WaveLab and Reproducible Research". In: *Wavelets and Statistics*. Ed. by Anestis Antoniadis and Georges Oppenheim. Vol. 103. Lecture notes in statistics. Springer-Verlag, pp. 55–81.
- Dye, Thomas S. (in press). "A Model-based Age Estimate for Polynesian Colonization of Hawai'i". In: *Archaeology in Oceania*.
- Gentleman, Robert and Duncan Temple Lang (Mar. 2007). "Statistical Analyses and Reproducible Research". In: *Journal of Computational and Graphical Statistics* 16.1, pp. 1–23.
- Gentleman, Robert et al. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Hothorn, T, L Held, and T Friede (2009). "Biometrical journal and reproducible research." In: *Biom J* 51.4, pp. 553–5.

Literature II

- Hothorn, Torsten and Friedrich Leisch (2011). “Case studies in reproducibility”. In: *Briefings in Bioinformatics* 12.3, pp. 288–300.
- Ioannidis, J P et al. (2009). “Repeatability of published microarray gene expression analyses.” In: *Nat Genet* 41.2, pp. 149–55.
- Knuth, Donald E. (1984). “Literate programming”. In: *The Computer Journal* 27, pp. 97–111.
- Leisch, Friedrich (2002). “Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis”. In: *Compstat 2002 — Proceedings in Computational Statistics*. Ed. by Wolfgang Härdle and Bernd Rönz. Physica Verlag, Heidelberg, pp. 575–580.
- Schulte, Eric and Dan Davison (2011). “Active Document with Org-Mode”. In: *Computing in Science Engineering* 13.3, pp. 66–73.
- Schwab, M., N. Karrenbach, and J. Claerbout (2000). “Making scientific computations reproducible”. In: *Computing in Science Engineering* 2.6, pp. 61–67.

Thank you for your attention.