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)
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)  

"potentially reproducible"  

89%  

**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

89%
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:

\[
a \leftarrow \text{rnorm}(100, \text{mean}=0.0) \\
b \leftarrow \text{rnorm}(100, \text{mean}=0.8)
\]

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

Descriptive Values

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.21</td>
<td>1.04</td>
</tr>
<tr>
<td>b</td>
<td>0.71</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Visualization

Analysis

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

* The Data...

* Descriptive Values...

* Visualization...

* Analysis...

* Using Noweb

* Latex Options
A Simple Example

The Data

We first generate some data:

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

Descriptive Values...

Visualization

Analysis

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

---

The Emacs Org-mode, Andreas Leha, 16 August 2011 12
A Simple Example

The Data
We first generate some data:

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

Descriptive Values

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.14</td>
<td>1.12</td>
</tr>
<tr>
<td>b</td>
<td>0.73</td>
<td>1.02</td>
</tr>
</tbody>
</table>

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

The Data
We first generate some data:
\[
a \leftarrow \text{rnorm (100, mean =0.0)}
\]
\[
b \leftarrow \text{rnorm (100, mean =0.8)}
\]

Descriptive Values

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.21</td>
<td>1.04</td>
</tr>
<tr>
<td>b</td>
<td>0.71</td>
<td>0.92</td>
</tr>
</tbody>
</table>

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

```r
#+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
```

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

---

**Demo.org 17% (35,0) Git:master (Org Fill)**
A Simple Example

The Data
We first generate some data:

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

Descriptive Values

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.21</td>
<td>1.04</td>
</tr>
<tr>
<td>b</td>
<td>0.71</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Visualization

```r
t.test(a, b)$p.value
```

A t-test shows that the means are significantly different (p-value: 0.000316729293949556).

Analysis

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

* Visualization...

* Analysis...

* Using Noweb

```org
#+begin_src R :noweb
<<generate_data>>
<<analyze_data>>
#+end_src

#+results:
  6.19585741319946e-07

* Latex Options

: :noexport:...
```
Export

```
<code>
+ .org
<text>
document
tangle
weave .tex/.pdf
.html
.odt
...
```
A Simple Example

The Data

We first **generate** some data:

\[
a \leftarrow \texttt{rnorm(100, mean=0.0)}\\
b \leftarrow \texttt{rnorm(100, mean=0.8)}
\]

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

Descriptive Values

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.08</td>
<td>0.99</td>
</tr>
<tr>
<td>b</td>
<td>0.84</td>
<td>1.08</td>
</tr>
</tbody>
</table>

Visualization

![Boxplot of generated data](image-url)
A Simple Example

1. The Data

We first generate some data:

```r
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

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.06</td>
<td>1.11</td>
</tr>
<tr>
<td>b</td>
<td>0.92</td>
<td>1.04</td>
</tr>
</tbody>
</table>

3. Visualization
Various Languages

<table>
<thead>
<tr>
<th>R</th>
<th>Python</th>
<th>Shell</th>
<th>Java</th>
</tr>
</thead>
</table>

\[\text{code}\] \[+\] \[.org\]

\[\text{document}\]

\[\text{tangle}\]

\[\text{weave}\]

\[\text{code}\]

\[\text{.tex/.pdf}\]

\[\text{.html}\]

\[\text{.odt}\]

\[\ldots\]
Various Languages

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

<table>
<thead>
<tr>
<th>lrr</th>
<th>mean</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>v7.6</td>
<td>139.2</td>
<td>2401.96</td>
</tr>
<tr>
<td>v7.7</td>
<td>40.48</td>
<td>81.34</td>
</tr>
</tbody>
</table>

A Wilcoxon test shows ‘no significant’ difference (p-value: 0.082).
Various Languages

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

```sh
#+begin_src sh :session none :exports none
  cd ${HOME}/local/emacs/org-mode
  git log --no-merges --oneline \n    --shortstat ${FROM}..${TO} | \n    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(-)   |
```

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


Thank you for your attention.