Using LaTeX and Markdown for Reproducible Research

Erik Beck

Emily Y. Li

US EPA
R User Group Workshop

13 August 2019
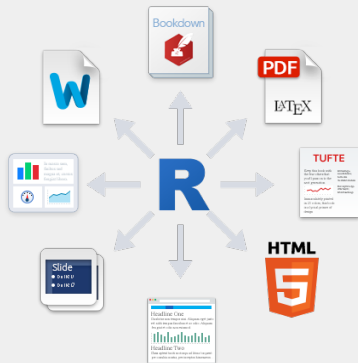
Table of Contents

Table of Contents

Using LaTeX and Markdown for Reproducible Research

Day	Time	Room
Tues, Aug 13	**8:30–12:00pm**	**C111C**

- This 1/2-day workshop will provide attendees with hands-on experience using the basics of Markdown and knitr.
- After attending this workshop, you will be able to use these tools to facilitate reproducible reports and research with R.

We will try to use our half day as effectively as possible.

Rough Agenda

#	Time	Topic
1	8:30	System checks & agenda
2	8:50	Reproducible Research
3	9:10	Intro to Markdown
4	9:30	Intro to LaTeX
5	9:40	Markdown & LaTeX examples
	10:10	*BREAK*
6	10:20	Dynamic documents with knitr
7	10:40	Markdown & LaTeX with R
8	11:40	Wrap-up & additional resources

TABLE OF CONTENTS

IS THERE A REPRODUCIBILITY CRISIS?

7%
Don't know

3%
No, there is no crisis

52%
Yes, a significant crisis

1,576
researchers surveyed

38%
Yes, a slight crisis

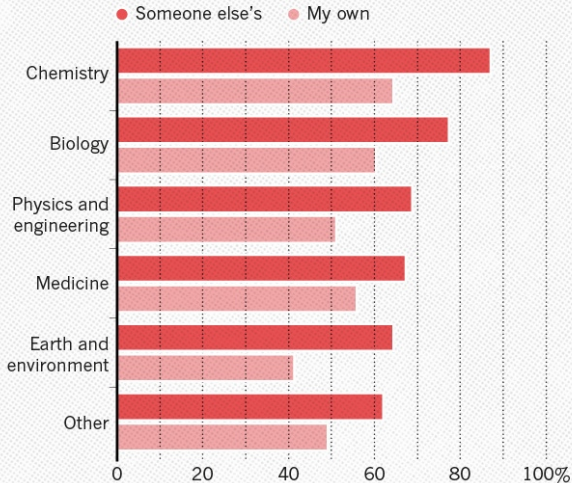©nature

Reproducible Research

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.

● Someone else's   ● My own

Chemistry
Biology
Physics and engineering
Medicine
Earth and environment
Other

0   20   40   60   80   100%

Results must be reproducible to be trustworthy.

*An article about computational science in a scientific publication is not the scholarship itself, it is merely the advertising of the scholarship. The actual scholarship is the complete software development environment **and the complete set of instructions which generated the figures**.*
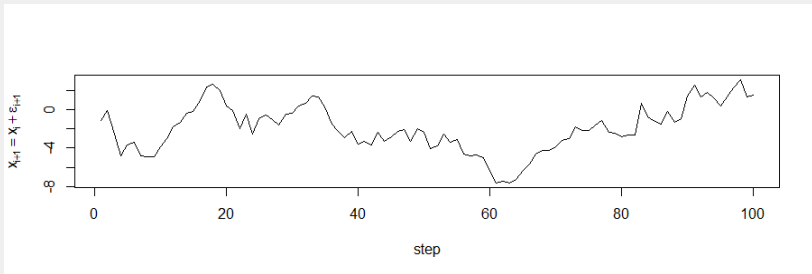
— 1995, David L. Donoho, professor of statistics at Stanford University

This chunk of R code produces a figure that illustrates a simulation of Brownian motion for 100 steps.

Try running this in RStudio

```
set.seed(1213) # for reproducibility
x <- cumsum(rnorm(100))
plot(x, type = "l",
     ylab = expression(x[i+1]==
                          x[i]+epsilon[{i+1}]),
     xlab = "step")
```

To put this into a document by hand, we would have to open RStudio, compile the code to draw the plot, save it as an image, then insert it into a document with \includegraphics{}.

Then what if we want to change the random seed in set.seed(), or the y-axis label?

- Instead of separating the results from the computation, we can put everything in one document.
- When we compile this document, the computation will be executed, giving us the results directly.
- Integrating code with narratives is not only easier, but also provides details needed for reproducibility.

Dynamic documents are easier than cut-and-paste

It is fairly common to see student homework and exercises among the countless user contributions on RPubs. Once students are trained, we may expect more reproducible scientific research in the future.

This table shows a subset of the *mtcars* dataset.

	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21.0	6	160	110	3.90	2.620
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875
Datsun 710	22.8	4	108	93	3.85	2.320
Hornet 4 Drive	21.4	6	258	110	3.08	3.215
Hornet Sportabout	18.7	8	360	175	3.15	3.440
Valiant	18.1	6	225	105	2.76	3.460

Have you ever included a table in your document? In LaTeX, the code would look like this…

Manual table in LaTeX

```
\begin{tabular}{ l r r r r r r }
     & mpg & cyl & disp & hp & drat & wt \\
   Mazda RX4 & 21.0 & 6 & 160 & 110 & 3.90 & 2.620
   Mazda RX4 Wag & 21.0 & 6 & 160 & 110 & 3.90 & 2.8
   Datsun 710 & 22.8 & 4 & 108 & 93 & 3.85 & 2.320
   Hornet 4 Drive & 21.4 & 6 & 258 & 110 & 3.08 & 3
   Hornet Sportabout & 18.7 & 8 & 360 & 175 & 3.15 8
   Valiant & 18.1 & 6 & 225 & 105 & 2.76 & 3.460 \\
\end{tabular}
```

Even if you maintain this table in Excel, there are manual steps
and human intervention required to update your document.

If you wanted to include this same table in your dynamic document, you would only need this code:

Table generated in R code

```
library(knitr)
kable(head(mtcars[, 1:6]))
```

If you ever update the mtcars dataframe, the table will update in the document on its own.

RPubs student example

Table of Contents

```
# First level header

Sup universe!

## Second level header

This will be **bold**, and *italic*.

− list item
− list item

You can write an ordered list:

1. item 1
1. item 2 # comment: this line will render as "2. ite
```

Markdown is an **easy-to-read, easy-to-write plain text format**.

Quickly lay out documents

Maximize focus on content

Minimize focus on formatting

Convert easily to a variety of formats

Always compatible for records, plain text format is durable

Limitations of Markdown

Markdown was primarily designed to be simple.
For more complicated typesetting, LaTeX may be preferred.

https://rmarkdown.rstudio.com/

- An **.Rmd file** is a Markdown file with R code chunks in it.
- Contains the **code** that a scientist needs to reproduce your work, along with the **narration** that a reader needs to understand it.
- Can export the finished report in a variety of formats, including HTML, PDF, or MS Word.

Workflow in Markdown

Using **RStudio**:

Open a new .Rmd file, which pre-populates with a template

Write a document by editing the template

Knit the document to create a report; use the knitr button or render() to knit

Preview output in IDE window

(optional) Publish to web server

Use output file that is saved alongside the .Rmd file

Helpful link:
https://rmarkdown.rstudio.com/lesson-2.html

Table of Contents

LaTeX is a tool for high-quality typesetting, designed for the production of technical and scientific documentation.

How do you pronounce "LaTeX"?

TeX is usually pronounced tech, making 'lah-teck, lah-'teck, and 'lay-teck the logical choices; but language is not always logical, so 'lay-'tecks is also possible.

— Leslie B. Lamport, original developer of LaTeX

LaTeX is widely used in academia for the publication of scientific documents in many fields, including mathematics, statistics, computer science, engineering, chemistry, physics, economics, and political science.

- TeX is fast, consistent, and stable.
- TeX engines have **excellent quality output**.
- Use it when...
  - ▸ You have many tables, figures, equations, cross-references, citations, or hyperlinks.
  - ▸ You want your final product to look exceptionally professional (like a resume).

— https://www.ctan.org/tex/

LaTeX: Minimal example

Here is a minimal example of a full document written in LaTeX.

```
\documentclass{article}
\title{A Minimal LaTeX Example}
\author{Emily Li}

\begin{document}
\maketitle

Yer a wizard, Harry.

\end{document}
```

TABLE OF CONTENTS

EXAMPLES & DEMOS

Let's try a little Markdown:

One Use Markdown to write a .html file with your name and favorite hobby.

Two Output the same file as a Word document.

Break for 10 minutes

Welcome back!

TABLE OF CONTENTS

Sweave and **knitr** can compile narrative documents with R code.

Sweave is included in base R, and specifically compiles R code integrated into LaTeX or LyX documents.

knitr compiles **any input language** (R, Python, SAS...) inside **any markup language** (LaTeX, Markdown, HTML...)

Sweave

Sweave has been a prominent, longstanding tool for dynamic documents since 2002.

- Combines the power of R with the production value of LaTeX
- Part of base R (as the `Sweave()` function)
  - From your R session: `Sweave("your_file.Rnw")`
  - From the command line: `R CMD Sweave your_file.Rnw`

However…

- Development has plateaued in recent years
- Not modular enough; extensions may become incompatible. Some packages are no longer synchronized.
- A PDF produced from LaTeX looks great, but is often not the format we need when collaborating.

knitr was largely motivated by Sweave, and designed to be easier to maintain and extend.

> *The design of knitr allows any input languages and **any output markup languages**.*

— Yihui Xie, creator of knitr

- Can compile dynamic documents in Markdown—which are more intuitive and human-readable.
- Works very well with Pandoc, so creating a Word document or OpenDocument format is just as easy as creating a PDF.

Check out `https://www.rdocumentation.org/packages/knitr` for more info on installation, motivation, usage, and functions.

Install knitr
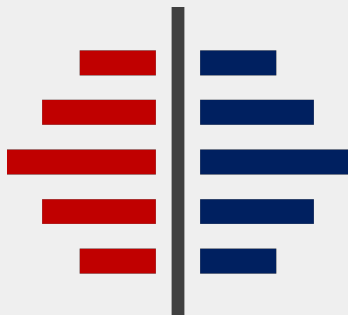
```
install.packages('knitr', dependencies = TRUE)
```

Go to the menu *Tools > Options > Sweave* and switch the default option for weaving (compiling) to **knitr**.

Visit https://yihui.name/knitr/faq/ for frequently asked questions.

Table of Contents

We'll look at a side-by-side comparison of
Markdown and LaTeX using the exact same analysis.

We will be using RStudio to edit and
compile the source code.

You can create either R-Markdown or R-LaTeX documents in RStudio.

- Files that mix R code and Markdown are **.Rmd** files
- Files that mix R code and LaTeX are **.Rnw** files

Why is the R-LaTeX file extension ".Rnw"?

"Rnw" stands for "R no web". NoWeb is an old but active simple system for mixing code and narratives.

Markdown and LaTeX side-by-side

R code and Markdown = **.Rmd**

- ```` ```{r} ```` opens a code chunk and ` ``` ` terminates it
- Inline R code is written in backticks ` ` `
- Chunk options are written before the closing brace } in the chunk header

R code and LaTeX = **.Rnw**

- `<<>>=` opens a code chunk and and @ terminates it
- Write chunk options in between `<<` and `>>=`

You had me at "keyboard shortcut"

In RStudio, for either *.Rnw* or *.Rmd*, quickly insert code chunks with the keyboard shortcut *Ctrl + Alt + I*.

DEMO *Rmd — input.Rmd*

It is also possible to generate a quick report from R script using
knitr's stitch() function.

Stitch a quick report

```
library(knitr)
stitch("your-script.R")
```

- stitch() provides a template so the user only feeds the
  template with one R script and knitr will compile the
  template to a report.
- Currently it has built-in templates for LaTeX (default), HTML,
  and Markdown.

See ?stitch for details.

Markdown (or LaTeX) can be combined with R using knitr to make longer reports that can be easily updated with a tidbit of new code.

Here we will show an example using data from US EPA and US Dept. of State on $PM_{2.5}$ air pollution.

- Particulate pollution is often in the news
- Especially lately with West Coast forest fires
- CAA primary air pollutant (NAAQS)

Switch over to Rstudio...

Let's try some examples together…

Air Example One

Use Markdown to create a report that performs summary statistics on a different variable than the concentration value presented.

Air Example Two

Use Markdown to display a plot of your choice from the air data.

TABLE OF CONTENTS

Additional Resources

Erik Beck
beck.erik@epa.gov

Emily Y. Li
li.emily@epa.gov

- https://www.rstudio.com/resources/cheatsheets/
- https://support.rstudio.com/hc/en-us/articles/200552056-Using-Sweave-and-knitr
- Documents source + output examples
  https://yihui.name/knitr/demos/
- Recently published student examples at Rpubs
  http://rpubs.com
- https://rmarkdown.rstudio.com/authoring_basics.html

THANKS FOR COMING!