Reproducible Research and R Markdown

Dr. Wesley S. Burr

2018-11-05

Steps Toward Reproducible Research

Making your research reproducible is the goal. It's not trivial!

This is not original ...

Full acknowledgement: I ripped off most of these slides from a statistician named Karl Broman at UW-Madison. He's the person who introduced me to the core concepts.

This has happened to me ...

```
Wesley, received the report, looks interesting. However, I see that
you only have results up to 2013 - why have you not used the latest
data up to 2016? Can you re-do the analysis?
  - (a colleague of mine at Health Canada)
```

And so has this …

The results in Table 1 don’t seem to correspond to those in Figure 2.

And this ...

How did I get those results again? What order did I run all this code in ... ?

And this ...

Where did this data come from?

And this ...

How did this processed data set get made? Where did those weird data records go?

And this ...

This code that used to work is now erroring ...

And this ...

"The attached is similar to the code we used."
(actual quote from another researcher)

Reproducible

versus

Replicable

Levels of Research (Code) Quality

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to **what** was done, is it clear **why** it was done? (e.g., why did you omit those six subjects?)
- Can the code be used for other data?
- Can you extend the code to do other things?

1. Everything should be scripted

If you'll do something (with data or code) once, you'll do it 1000 times.

2. Organize your data and code

```
File organization and naming are powerful weapons against chaos.
- Dr. Jenny Bryan (UBC and RStudio)
```

2. Organize your data and code

```
Your closest collaborator is you six months ago, but you don’t
reply to emails.
- Dr. Mark Holder (UBC-Okanagan)
```

```
... looking at code from six months ago which appears to have
been written  by a stranger, possibly an insane stranger ...
- Dr. Hadley Wickham, Chief Scientist, RStudio (2016)
```

2. Organize your data and code

Directory structure! Coherent and straightforward process within scripts.

```
[wburr@cascade Code]$ ls
analyzeSims.R      genFig1_2_S1.R     genFigS2.R             packages
chicagoAnalysis.R  genFig1_2_S1.Rout  genFigS2.Rout          README.md
data               genFig3.R          installCRANpackages.R  sim_fullModels.R
figures            genFig4.R          Makefile               tables
[wburr@cascade Code]$
```

(and this is a not particularly good example - I've learned a lot since 2013)

3. Automate the process (GNU Make)

Not strictly necessary for science, but invaluable for **reproduciblity**. We can talk more about this in the future ...

4. Turn scripts into reproducible reports

This is what we're going to show you today!