Introduction to R and R Markdown for Reproducible Research

John M. Drake

Odum School of Ecology
University of Georgia
Athens, Georgia USA 30602-2202

jdrake@uga.edu

June 11, 2015

Population Biology
of Infectious Diseases
REU site @ UGA

Table of Contents

Three R's: Replication, repeatability, and reproducibility

Scientific knowledge aims to be general *in some sense*. "Three R's" underwrite this generality:

1. *Replication* concerns the number of data points (observations, study subjects, etc.) and establishes the generality of the observed finding to a study population.
2. *Repeatability* concerns the ability to arrive at the same findings when a study is repeated and establishes the generality of the observed finding to other study populations or systems.
3. *Reproducibility* concerns the reliability of the logic that leads from data to conclusions – that is, the *data analysis*. It would seem that reproducibility is an essential ingredient of scientific knowledge. But, as data workflow become more and more complicated they also bring more subjective decision-making by the data analyst, more computations, and more opportunities for error.

What is reproducible research?

Johns Hopkins Bloomberg School of Public Health:

- Reproducible Research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them.[1]

[1] https://www.coursera.org/course/repdata

Why should research be made reproducible?

Four public virtues and one private one

- Enables readers to develop a complete and exact understanding of methods
- Enables error detection and correction
- Facilitates critique
- Enables extension and advancement
- Saves time (private)

How can research be made reproducible?

Three concepts

- Workflow/Dataflow
- Literate statistical programming
- Science archive

Workflow/Dataflow

Workflow...

- is the idea that data analysis may be viewed as a repeatable pattern of computational activities. If these activities are truly repeatable and truly computational, then they may be encoded in an algorithm (programming). This principle of repeatable computation is the key to reproducible research. In reproducible research, a computer program is written to perform an analysis.

Dataflow...

- is the idea that when a variable changes, downstream computations affected by that variable should change as well.

Literate statistical programming

Literate programming...

- is an approach to programming in which computer codes are interspersed with explanations in natural language.

Literate *statistical* programming...

- is the application of the literate programming idea to statistical codes.

Science archive

For the public virtues of reproducible research to be realized requires *access* to...

- Data
- Computer codes
- Explanation

These can be bundled together and archived in a public repositor such as the Dryad Digital Repository[2]

[2] http://datadryad.org/

Arithmetic

Key commands to issue from the command line:

- assigment using <-, e.g., a<-5
- arithmetical operations +, -, *, / e.g., a+10
- "combine" values using c, e.g., `data <- c(2, 6, 16, 18)`

```
[1] 20
```

Exercise

You know that R_0 for a simple *SIR* epidemic is given by $R_0 = \beta/\gamma$. Calculate R_0 for an epidemic where $\beta = 1.2$ and $\gamma = 0.6$.