

GITting started with reproducibility: An introduction to `git` and `knitr`

Biostatistics Student Association Computing Workshop

Nick Seewald

Department of Statistics
University of Michigan

January 29, 2016

Why do I care about reproducibility?

Reproducible research is a hallmark of the scientific method, but we're pretty bad at it.

In 2012, a researcher then at the biotechnology company Amgen wrote in Nature that when his team tried to reproduce 53 landmark cancer studies, they could replicate just six. And according to a news report in Nature, a project aiming to reproduce the findings of 100 psychology papers has managed to replicate results for only 39 of them (the project's findings are still under peer review).

"What Science Can Tell Us About Bad Science", *The Atlantic*, September 2015. <http://www.theatlantic.com/magazine/archive/2015/09/a-scientific-look-at-bad-science/399371/>

But I'm a Biostatistician!

- ▶ Reproducibility is important in both science AND statistics!
- ▶ As statisticians, we need to be able to reproduce our results on the same data set
 - ▶ This means we have to write reports in a way that minimizes error and write code so that we can get the same results years later.

Agenda

1. Git: A “version control” tool used for collaborating and maintaining different versions of a file, typically for code.
 - ▶ Great for collaborating, or just saving your own ass.
 - ▶ Often used in conjunction with *GitHub*, an online repository storage service.
2. knitr: An R package that lets you create documents containing R code and output.
 - ▶ Keep everything you need to generate a report (e.g., for research, homework, or 699) in one place!
 - ▶ My favorite part: Update code without having to re-create tables! (This is where errors creep in!)

A Brief Warning



Source: <https://xkcd.com/1597/>

Setup

Create a GitHub account, then download either Git or GitHub Desktop.

- ▶ Pure Git (i.e., just command-line tools):
<https://git-scm.com/downloads>
- ▶ GitHub Desktop (GUI & command-line tools):
<https://desktop.github.com/>

clone a Repository

- ▶ To create a local copy of an existing git repository, use `git clone [url] [directory-name]`.
- ▶ In a terminal (Mac, Linux) or the Git Shell (Windows), navigate to the folder you want to clone the repository into.
- ▶ **Exercise:** Clone my bsa-computing repository onto your computer. The URL is `https://github.com/nseewald1/bsa-computing.git`

Agenda

1. Git: A “version control” tool used for collaborating and maintaining different versions of a file, typically for code.
 - ▶ Great for collaborating, or just saving your own ass.
 - ▶ Often used in conjunction with *GitHub*, an online repository storage service.
2. knitr: An R package that lets you create documents containing R code and output.
 - ▶ Keep everything you need to generate a report (e.g., for research, homework, or 699) in one place!
 - ▶ My favorite part: Update code without having to re-create tables! (This is where errors creep in!)

What is knitr?

- ▶ `knitr` lets you embed code and output from R into \LaTeX , HTML, RMarkdown, etc.