# Introduction to Reproducible Research

*Jonathan GIlligan*

*August 19, 2017*

## Why Do We Need Reproducible Research?

In the spring of 2012, Bruno Iksil, a securities trader at the investment bank JPMorgan, Chase, & Company who was also known by the nickname "The London Whale" for his aggressive trades, made a series of costly mistakes that cost JPM-Chase $6.2 billion. Iksil was attempting to manage the financial risk of a portfolio of investments. However, an analyst on Iksil's team had calculated the volatility (a measure of financial riskiness) of his portfolio using an Excel spreadsheet and made a subtle error in a mathematical formula, dividing by the sum of two numbers instead of the average. That error caused Iksill to underestimate risk by a factor of 2, and thus to expose JPM-Chase to far more risk than he realized with his enormous trades.

Two years earlier, Carmen Reinhart and Kenneth Rogoff, two highly respected academic economists published an influential research paper on the effect of government debt on economic growth. This paper concluded that when government debt exceeds 90% of GDP, the country's economic growth is likely to abruptly come to a halt, and even slide into recession. This paper was used to justify harsh austerity measures throughout Europe, where nations were struggling to recover from the 2008 global economic meltdown, and was cited by Paul Ryan as highlighting the importance of dramatically cutting federal spending in the U.S.

Thomas Herndon, a graduate student in economics at the University of Massachusetts, was skeptical about this research, but Reinhart and Rogoff's paper did not explain all the details of their data and analysis. Finally, in 2013, Reinhart and Rogoff gave Herndon copies of the spreadsheets they had used in their analysis. Herndon found three glaring errors in the spreadsheet, and after he fixed the errors, there was no sudden slowdown of economic growth.

Such errors are not unique to economic research. In the past two years, errors in spreadsheet formulas led to the retraction of papers in prominent journals of environmental science, medicine, and biology.

Spreadsheets are not the only source of major errors in scientific publications. Poor statistical practices have led to what has come to be called a "crisis of replication" in psychology and medicine and concern that many published scientific results are incorrect.

Of particular concern is the fact that major problems are being discovered in clinical medical research. Once recent review of 5,000 papers in eight top medical journals found that almost 100 had major inaccuracies.

For the most part, science works. Advances in all fields of science have led to deep understanding of nature, and have led to technological breakthroughs that drive our economy, enable us to live much longer and healthier lives, and otherwise improve the quality of our lives.

Nonetheless, even if only a few percent of major scientific research papers are wrong, this has potential to mislead us about which medicines or medical procedures are safe and which are dangerous, about which government policies are likely to be effective, and in the private sector, can lead companies to make financially disastrous mistakes.

Two important principles in sceince, which should prevent these errors, are that research should be *transparent* and *reproducible*: Research reports should describe the procedures clearly and in enough detail that other scientists know exactly what was done. And scientists who repeat the research procedures, as described in the reports, should find similar results, within the limits of experimental uncertainty.

However, as the anecdotes above, and hundreds of similar reports of problems in research reveal, too often even well-meaning scientists fall short of providing enough detail about their methods for other scientists to understand their work and catch errors, and it is often difficult to truly reproduce previously published research.

To address these problems, the scientific commuinity is increasingly embracing the principles of what has come to be called **reproducible research**.

Federal funding agencies, scientific journals, and scientific societies now call for authors to reveal all the details of their experiments and analysis, and must share the data and computer codes they used to perform the analyses described in their publications.

# What is Reproducible Research?

*Reproducible research* seeks to make scientific research completely reproducible by documenting every decision a researcher made in the course of collecting and analyzing data. At the simplest level, this would mean that when a scientist submits a paper to a research journal, she would include all the data and a clear description of the analysis.

However, a written description of the analysis process might inadvertantly omit crucial steps, or the researcher might describe what she thought she did, but might have made errors in her actual analysis.

In the example of Reinhart and Rogoff's paper on debt and economic growth, the two economists described what they thought they had done, but they were unaware that their spreadsheet contained errors. For three years after they published the paper, the errors remained buried in their spreadsheets but other economists only knew the written descriptions of the analysis that appeared in their paper and could not examine the spreadsheet for themselves.

Thus, reproducible research calls for researchers to share not only their data, but also any spreadsheets, computer programs, or scripts they used to perform the analysis. This will allow other researchers to catch errors where the actual analysis procedure does not match the desccription in the published report, just as Thomas Herndon was able to do when he obtained Reinhart and Rogoff's spreadsheets.

## Scripts versus Spreadsheets

In principle, this should suffice, but in practice it turns out that auditing a spreadsheet is very difficult. When you open a spreadsheet in Excel, you see a grid of text and numbers, but the formulas used in calculating the values of certain cells from other cells are largely invisible and it is difficult to read and audit every formula in a spreadsheet that contains thousands of cells.

Thus, the scientific community has become increasingly mistrustful of spreadsheets and prefers data analysis tools that use scripts to conduct the analysis. Scripts (basically, short computer programs) are written in a textual form that is straightforward for a knowledgeable person to read and understand. Consistency checks to catch errors are much easier to implement in scripts than in spreadsheets.

In the labs for EES 3310 and 5310, we will be using scripts for the R statistical analysis program to analyze climate data and the output of climate models. In the course of the semester, you will become increasingly familiar with the ideas of using analysis scripts to promote reproducible research.

## From the Analysis to the Manuscript

Even when analysis is performed correctly, it can be difficult to transcribe every number correctly into the manuscript of a research report. When I (Professor Gilligan) was in graduate school, one of my professors told me a cautionary tale from early in his career: Three prominent physicists were attempting a very difficult calculation in quantum electrodynamics. To guard against errors, each of the three performed the calculation independently and then they compared their results. They were very excited to discover that their calculations agreed perfectly. One of them went to add up the different terms from the calculations and type the result into the manuscript of a paper they rushed into print to announce and share their accomplishment.

Somewhere in the process, he transcribed a number incorrectly, so the published result was incorrect and the embarrassing error was not discovered until some time after it appeared in print.

Consider, too, what happens if a research report is almost complete and the researchers discover an error in their analysis scripts or in their raw data. After they make the correction, they must adjust every number in their final manuscript. This introduces additional risks of either mistyping numbers or of missing a number that needs to be changed.

With modern computing tools, it has become easy to integrate the analysis with the final report.

In the EES 3310 and 5310 laboratories, we will use a tool called RMarkdown, which allows you to combine the text of your lab reports with the scripts that perform your data analysis and generate the figures and tables for your report. Thus, any time you change a number in your data or change a line of code in your analysis script, the computer can automatically regenerate your report to update all of the numbers and figures accordingly.

# Elements of Reproducible Research

This section is adapted from the "Introduction to Reproducible Research" by the R Open Science Project, http://ropensci.github.io/reproducibility-guide/sections/introduction/

## Kinds of Reproducibility

Reproducibility means different things in different scientific fields. One big distinction is between computational versus observational or empirical aspects of research.

- Computational reproducibility provides detailed information on exactly how computations (either calculations or simulations using computer models) were performed, and making it possible for others to exactly reproduce those computations or calculations. This includes providing source code for programs and scripts written by the researcher, together with detailed specifications of the software (including the specific versions used), and the hardware used (running the same software on different computer hardware can sometimes give different results). R has a function that will automatically report the computing hardware and software, so it will be trivial to add this to the end of all your laboratory reports.
- Empirical reproducibilty provides detailed information about laboratory or field procedures that were used to acquire empirical data used in the analysis. In practice, this is often accomplished by providing the raw data together with details about how it was collected.

In this laboratory, we will focus entirely on *computational reproducibility*. We will not address all the details of computational reproducibility in this laboratory, but we will focus on three important aspects:

- **Literate computing and authoring:** Litrerate computing refers to mixing computer code with narrative description of what the code is doing. Stanford computer science professor Donald Knuth invented literate programming based on his experience with large software projects, where he found that if he wrote clear narrative descriptions of what his programming code was doing, he made fewer errors, and could find and correct those errors more quickly.

  In this laboratory, we will use RMarkdown and RStudio to apply literate computing and authoring in laboratory activities and writing reports.

- **Automation:** Many of you may be used to so-called "point-and-click" software tools for statistical analysis. Examples include Excel, SPSS, and Stata. These tools let you perform analysis by reading in your data and then highlighting data with a mouse and selecting menu options. This approach makes it easy to get started with these software packages when you are a beginner, but make it difficult to track exactly what you did in your analysis, so when it comes time to write up your report, you may not

remember exactly what you did, and in what order you did it. Automating your analysis using scripts means that your script contains the complete information about everything you did.

Some programs, such as Stata, allow you to record your analysis and export a script (a `.do` file) that will allow you or others to reproduce your analysis. This is a valid form of reproducible research, but it is not the one we will use in this course.

Automation is also important if you have to do the same operation on many different data sets. What seemed easy when you just had to create one graph or analysis can quickly become tedious as you have to drag the mouse and click on the same menu entries over and over again on a dozen different data sets. Automating your analysis with scripts makes it easy to run the same script on each of your different data sets.

- **Revision Control:** As you edit both your text and the R scripts you use for your analysis, it is valuable to be able to keep track of changes. For instance, if your analysis is working well, and then you edit something and it stops working it is useful to be able to go back and look at what changed between the time when it was working and when it stopped working.

  Revision control systems allow you to easily keep track of changes you make to your files.

  Another important use of revision control is not relevant to this laboratory section, but applies to larger research projects. I have computational models that are constantly under development and I publish papers based on them. Suppose that another scientist has a question about a paper that I wrote two years ago, but I have made many changes to the model since then. How can I go back and recreate the version of the model that I used for that paper? Revision control systems make this very easy.

  Finally, revision control systems are very useful for team projects because they allow a team of many researchers to coordinate their activities when they are all editing files (computer code and text) for a project at the same time.

  We will be using a revision control tool called `git`. There is a web site called github.com, which allows people to share projects. My students and I use github to release software that we develop in our research, and there is an educational site connected to github, which we will use for laboratory assignments in EES 3310 and 5310.

  Everyone should sign up for a free student account on github.

## Software Tools

All the software you will need is installed on the computers in our laboratory classroom (SC 2200, right next to the Science and Engineering Library). If you want to install it on your personal computer, all the software we use is free and open source.

- R https://cran.rstudio.com/ Available for Windows, Mac OS X, and Linux
- RStudio is available in many different options. What you want is the Open Source version of RStudio Desktop, which you can download from https://www.rstudio.com/products/rstudio/download/#download. Be sure to install R before you install RStudio.

  After you install RStudio, you will almost certainly want to install several optional packages for R that we will use extensively for the labs. From the "packages" window in RStudio, click on "Install" and install the following packages: "tidyverse", "knitr", "rmarkdown", "xts", "lme4", "ggalt", "GGAlly", "ggExtra", "ggmap", "ggspectra", "ggspatial", "ggspectra", "ggstance", "gtable", "hrbrthemes", and "pacman"
- Git for revision control. You have three options:
  - You can download Git from https://git-scm.com/

- There is a very popular free git client called "Source Tree" that has some very nice graphical utilities that let you use git from Windows Explorer or Mac Finder. You can get Source Tree from https://www.sourcetreeapp.com/ for Mac or Windows. You will need to sign up for a free account at Atlassian to install it.
- A third option is Git Kraken, which you can get from https://www.gitkraken.com/ Git Kraken is very popular, and is free for educational, personal, and other non-commercial use. Git Kraken is available for Windows, Mac, and Linux.

After you install Git, it is important to run two commands:

- Open a git command line:
  - On Windows, open the program menu, go to "Git" and click on "Git Bash"
  - On a Mac, open a terminal window
- Run the two following commands:
  - `git config --global user.name "Your Name"` (using your own name instead of "Your Name")
  - `git config --global user.email "your.email.address@vanderbilt.edu"` (using your own email address)

Git uses information to keep track of who makes changes to a file. If you are editing a file on your computer and a friend is editing it on her computer, git uses this user information to keep track of who made each change. Then when you and your friend merge your changes, git will be able to tell you which of you edited what.

### Optional Tools

If you want to produce PDF files from RMarkdown, you will need to install the LaTeX software.

If you have a Windows computer, I recommend MikTeX, which you can get from https://miktex.org/download

If you have a Mac, I recommend the MacTeX distribution, which you can get from http://www.tug.org/mactex/

On Linux, you should be able to install LaTeX using the package manager for your Linux distribution (e.g., `sudo apt-get install texlive-full` on Debian, Ubuntu or Mint; and `sudo yum install texlive` or `sudo dnf install texlive` on Fedora, CentOS, or other RedHat-based distributions)

## Walking the Walk

Over the last five years I have become increasingly convinced that reproducible research methods are both more efficient and also lead to higher quality reseearch. In my own research, I use the methods and many of the tools that we will use in this laboratory.

## Further Reading

If you are interested in learning more about reproducible research, I would recommend the following:

- Christopher Gandrud, *Reproducible Research with R and RStudio* (Second Edition) (CRC Press/Chapman & Hall, 2015). Gandrud is an economist and political scientist, who pioneered a lot of the methods that I use for reproducible research as part of his Ph.D. dissertation. This book is a comprehensive how-to guide to reproducible research, and all of the files necessary to reproduce the book are available online at https://github.com/christophergandrud/Rep-Res-Book

- The R Open Science Project, *Reproducibility in Science: A Guide to Enhancing Reproducibility in Scientific Results and Writing* http://ropensci.github.io/reproducibility-guide/