Reproducibility made simple

Automating reproducible research workflows

Aaron Peikert

2020-05-21

Contents

Abstract

This is a short summary.

1 Theoretical Considerations

Claerbout & Karrenbach (1992) define reproducibility as the ability to gain the same results, from the same dataset. Conversely, they call a result replicable if one draws the same conclusion from a new dataset. This thesis concerns itself with the former, providing researchers with an accessible analysis workflow, that is virtually guaranteed to reproduce across time and devices. The scientific community agrees that their work should be ideally reproducible. Indeed it may be hard to find a researcher who distrusts a result because it is reproducible; to the contrary, many feel it is "good scientific practice" to ensure it is ("Reducing our irreproducibility," 2013; Deutsche Forschungsgemeinschaft, 2019; Epskamp, 2019). Several reasons, practical and meta-scientific, justify this consensus of reproducibility as a minimal standard of Science.

Reproducibility makes researchers life more productive in two ways: The act of reproduction provides, at the most basic level, an opportunity to spot errors, helping the researchers who originally produced them. At the same time, other researchers may benefit from reusing materials from an analysis they reproduced.

Beyond these two purely pragmatic reasons, reproduction is crucial, depending on the philosophical view of Science one subscribes to, because it allows independent validation and enables replication. Philosophers of Science characterise Science by a shared method of determining if a statement about the world is "true" (Andersen & Hepburn, 2016) or more broadly evaluating the statements verisimilitude (Gilbert, 1991; Meehl, 1990; Popper, 1962; Tichỳ, 1976). If this method is for experts to agree on the assumptions and deduce some truth, reproducibility is hardly nec-

essary. On the other hand, it gains importance if one induces facts by carefully observing the world. The decisive difference is that the former gains credibility through the authority of the experts, while the latter is trustworthy because anyone may verify it. Accepting induction as a scientific method hence hinges on the verifiability by others. Some have even argued that such democratisation of Science is what fueled the scientific revolution (Heilbron, 2004, Scientific Revolution). The scientific revolution had the experiment as an agreed-upon method to observe the reality and a much later revolution provides statistical modelling (Rodgers, 2010) as a means to induction. This consensus, about how to observe and how to induce, gives modern scientific enterprises much of its credibility. Two reasons justify why we must assume reproducibility as a scientific standard if we accept induction as a scientific method: First, it enables independent verification of the process of induction, and second, it dramatically simplifies replication as a means to verify the induced truths.

However, neither the practical reasons that results may be less error-prone and more reusable nor the meta-scientific grounds that the process of induction and the induced facts are more straightforward to verify, if reproducible, follow strictly from the definition on reproducibility provided by Claerbout & Karrenbach (1992) given above. A simple thought experiment illustrates this shortcoming. Imagine a binary program that is perfectly reproducible; hence upon input of the same dataset, it fills a scientific manuscript with the same numbers at the right places. Furthermore, assume this hypothetical program may never hold if the data changes. Does the predicate "reproducible" in situation reduce the number of mistakes or enables reuse? Unlikely. Or could one audit it and use it in replication? Hardly. This admittedly constructed case of a reproducible black box shows us: we are not interested in reproducibility, we are interested in its side effects.

Spoiling its elegant simplicity, I change the definition by Claerbout & Karrenbach (1992) to address this issue, by further demanding that reproducibility must facilitate replication. Hence, I would call a result only then reproducible if the results remain unchanged if the data does, and

it furthermore helps other researchers to replicate the results if they attempt to do so. With such a notion, the only valid cause of reproducibility is transparency. Only if it is clear how the data relates to its results, both reproducibility and replication get promoted. It follows that something is no longer either reproducible or not, but there are shades, because a research product may promote replication to varying degrees. Note, that a scientific result can facilitate replication without anyone ever attempting to replicate it, e.g. by educating other researches about the analyses method, being openly accessible and providing reusable components.

Hence reproducibility has a technical side, ensuring the same results, and a non-technical side, facilitating understanding. The former relates to the practical advantages while the latter serves the metascientific purposes of reproducibility. An important caveat of the technical aspect is that generating the same results from the same data should be possible regardless of time and machine. Following a reproducible analysis should be:

1. understandable by other researchers
2. transferable across machines
3. conserved through time.

This much more demanding standard of reproducibility gains justification by two recent developments in the social sciences in general and psychology in particular: the emergence of a "replication crises" (Ioannidis, 2005) and the rise of "machine learning" (Jordan & Mitchell, 2015) as a scientific tool. Both trends link to the use of statistical modelling on which the social sciences became reliant for testing and developing their theories (Gigerenzer et al., 2004; Meehl, 1978). It turns out, if one fits the same statistical model as published on newly gathered data, one fails to achieve the same results as published more often than not (Open Science Collaboration, 2015). Such failure to replicate findings previously believed to be robust has amounted to a level some social scientists term a crisis. They put forth various causes and remedies to this crisis. Most remedies share a common theme: transparency. Some call for Bayesian statistics (Maxwell et al., 2015), as it makes assumptions more explicit, or demand preregistration (Nosek et al., 2018) as a

means to clarify how to analyse the data, beforehand and publicly, others require the researchers to publish their data (Boulton et al., 2012). Similar calls for transparency, as a response to the replication crises, have formed the open science movement which stresses the necessity of six principles (Kraker et al., 2011):

- Open Access
- Open Data
- Open Source
- Open Methodology
- Open Peer Review
- Open Educational Resources

I argue that a research product resting on these pillars optimally facilitates replication and hence satisfies the highest standard of reproducibility. If everyone has access to a scientific product and its data along with the source code, leading them to understand the methodology and thus enabling them to criticise the result and educate themself, one is in the best position to replicate it. Hence, any one's ability to reproduce such result gives a tangible affirmation of its usefulness to the scientific community.

However, reproducibility is no hurdle when anyone can perform the calculations needed with a pocket calculator; however, the more and more frequent use of computer-intensive methods renders such expectation questionable. The use of machine learning techniques, once enabled by the computer taking over strenuous works, now impedes our quest for reproducibility. More massive amounts of more complicated computer code than ever create room for errors and misunderstandings, leading the machine learning community to believe that they face a reproducibility crisis themselves (Hutson, 2018). Yet, I am far from calling for abstinence from machine learning, just because it complicates reproduction, but want to emphasise the need for solutions that allow anyone to reproduce even the most sophisticated analysis.

Peikert & Brandmaier (2019) put forth an analysis workflow which provides this accessibility for everyone to reproduce any analysis. However, they fail to provide the same level of convenience for the researcher who created an analysis in the first place. Setting up the workflow eats up a considerable amount of the researcher's time, which they may better spend at advancing research. This additional effort offsets the increase in productivity, promised by reproducibility, which I regard as most significant in the workflows adoption. Persuading researchers, who find the meta-scientific argumentation noble but impractical, do not care about it or oppose it, requires concrete, practical benefits. Luckily, most of this setup process may be automated, letting the researcher enjoy the workflows advantages while decreasing the efforts necessary to achieve them. Providing an easier to use and more accessible version of the analysis workflow by Peikert & Brandmaier (2019) is the goal of this thesis and the herein presented `repro`-package for the R programming language (Peikert, 2020).

2 Technical Solutions

This section summarises the workflow proposed by Peikert & Brandmaier (2019; see also The Turing Way Community et al., 2019 for a very similar approach). They argue that to ensure reproducibility, publically sharing code is not enough. Instead, reproducibility has to rest on five pillars:

1. **file management** a folder containing all files, referring to each other using relative paths
2. **literate programming** a central dynamic document, that relates code to thought
3. **version control** a system in place that manages revisions of all files over time
4. **dependency management** a formal description of how files relate to each other
5. **containerization** an exact specification of the computational environment

These pillars stipulate the relations between thought, code and data with their change over time and environment and hence reach all requirements of reproducibility through being:

1. understandable to other researchers,
2. transferable across machines,
3. conserved through time.

While comprehensibility to the scientific community, is probably the most crucial goal, it is the most difficult to achieve. That is because as a non-technical requirement, no set of rules may assure its fulfilment (though clear writing[1] and clean code[2] certainly help). Transfer and conservation, on the other hand, are problems with technical solutions. Peikert & Brandmaier (2019) propose to use a combination of RMarkdown, Git, Make, and Docker, because they are the most popular solutions for users of the R programming language (R Core Team, 2020). However, they stress that any combination of tools is suitable as long as it facilitates the above pillars.

Each of the following sections first raises a challenge for reproducibility, then outlines a conceptual remedy along with a concrete tool and concludes how they relate to the package `repro`. The relation of these tools with `repro` is then expanded in the next chapter.

2.1 File Organisation

File organization has to meet two challenges. First, the structure needs to be understandable by others, and second, it needs to be self-contained so that it can be moved to another machine.

Adhering to conventions help other people understand how files are organized. For example, the filename `R/reshape.R` follows both standard naming conventions (all lowercase, ends with .R, placed within the R directory) and is meaningful. Contrarily, `myScripts/munge_Data.r`

[1] Williams (2017) provides some excellent principles for writing clearly.
[2] Martin (2011) proposes a coding paradigm that found widespread use because of its focus on understandability.

is probably a lot harder to understand and remember for most R-users.

Following two guidelines makes the file structure self-contained:

1. Everything is in one folder.
2. Every path is relative to that folder.

This simple concept of a self-contained folder is facilitated by two R specific tools, RStudio projects and the `here` package (Müller, 2017). The former frees the user from changing the working directory manually; the latter infers absolute paths from relative ones. However, unlike the native R solution, this inference is consistent across operating systems and scripts and RMarkdowns.

The `repro` package comes with a template for an RStudio Project, which sets up a file structure that follows best practices and conventions.

2.2 Dynamic Document Generation

Even when following the most logical structure, it may be difficult for a reader of a scientific document to understand how the content of the document relates to the alongside published code. Providing a direct link, dynamic document generation allows interspersing text with code and its results, producing one human-readable document. The key feature is that every time such a document is rerun, the results are reproduced dynamically. This functionality eliminates errors due to copying and pasting results from statistical software to a text processor. This mistake may be all too common; Nuijten et al. (2016) reports that 50% of papers from the psychological sciences contain an error that may be prevented.

RMarkdown does provide a convenient framework to write such dynamic documents and render them as a wide range of output formats[3]. In an RMarkdown, three parts can be distinguished:

[3]The document you are viewing also results from a collection of RMarkdowns availible as website, PDF and E-book

- one specifying its output and metadata,
- one containing code, and
- one with descriptive text.

Each part uses its own language, all of them designed with ease of use and readability in mind. The section containing the output format and other metadata alongside is written in YAML (see the example below). This specification is located at the top, separated by three dashes at the beginning and end of the section. (R-)Code executing an analysis can be placed in a distinct chunk or inline within the text. The former has three backticks on their own line signifying beginning and end. The later is quoted in a pair single backticks. Examples of both methods can be found below. Text, which is not fenced by either three dashes or backticks, is interpreted as literal text written in the Markup language "Markdown". Markdown allows annotating text to signify formattings such as bold, italic, links and the inclusion of images.

The following section shows examples of metadata, code and text, specified as above described, forming a minimal example of an RMarkdown (adapted source code from Xie et al. (2019)/CC BY-NC-SA 4.0):

```
---
title: "Hello R Markdown"
author: "Ross Ihaka & Robert Gentleman"
date: "1997-04-23"
output: pdf_document
---
```

````
This is a paragraph in an R Markdown document.

Below is a code chunk:

```{r}
````
````

```
fit = lm(dist ~ speed, data = cars)
b   = coef(fit)
plot(cars)
abline(fit)
```

The slope of the regression is `r b[1]`.

Resulting in this document:

[fixme]

The `repro` package extends the yaml metadata to incorporate Dependency Management and Containarization into the process of dynamic document creation.

2.3 Version Control

Text, code and results of a scientific document are refined in cycles of many revisions to accommodate highest standards. As changes accumulate, different versions do too, posing a problem for reproducibility as it may be challenging to find out which version of code relates to the final product. One may argue that in the typical publication process, the final product is apparent: the published paper, so only one version is relevant. However, reproducibility may be crucial even before publication as part of the peer-review process. Also, recent trends in the publication process like preprints, open review, registered reports and post-publication review, blur the lines between published and unpublished.

To organize different versions as changes accumulate across the phases of a project across machines and users is a well-known challenge in software development. This challenge is met by

a high degree of automation that keeps track of different versions and has advanced facilities to compare and merge them.

One such version control software is Git. Git tracks versions of the project folder by taking snapshots of a given state called commits. Each commit has a unique id, called a hash, a short description of the changes made, called commit message and a link to the previous commit. This linking creates a "pedigree" of versions where it is easy to see how things have evolved. Going back in time to a specific version only requires to know the hash of the commit. To mark commits as special milestones, they can be tagged, e.g. as preregistration, preprint, submission or publication.

While mastering Git requires some experience, most of the time, only four commands are needed, which may be accessed through RStudio's Git interface:

git add take a snapshot of the given file
git commit create a commit of all added files
git push upload recent commits to a server
git pull download and integrate recent commits from the server

While a few other commands are necessary to set up Git in a given project directory, this work is done by the `repro`-package.

2.4 Dependency Management

In an analysis, the results depend on code which in turn depends on data. However, seldomly the data is analyzed as is, but some code is dedicated to preparing it. Most likely, each analysis needs a slightly different version of the data. An analysis of missingness requires the missings to be retained, but some statistical models do not allow that. Or the modelling software requires data to be differently shaped, then the plotting library. It is also often the case that one analysis

is based on the output of another and so forth. As these relations can become quite complicated, it is necessary to make them explicit to avoid confusion. Dependency management provides a formalism that describes how files depend on other files. More specifically, it provides an automated way to create files from other files, e.g. it automatically generates a cleaned version of the data, by relying on a cleaning script and the raw data. Such relations may be layered; hence, if a plot requires this cleaned dataset, first the cleaned dataset and then the plot is generated automatically. Such structure allows saving considerable computing time, as dependencies are not generated again if they already exist, but only if one of their dependencies has changed. In this example, upon recreation of the plot, the cleaned dataset is not generated as long as the cleaning script and the raw data remain unchanged. Such intelligent behaviour is most useful when the preprocessing requires a lot of computing time as is typical in neuroimaging or machine learning.

Make is a tool for dependency management, while originally designed for the compilation of programs, it is now increasingly recognized as a tool for reproducibility. It allows for all features above and more as it is an own programming language.

However, the repro package provides a much-simplified interface to the essential features, eschewing the need to learn yet another language.

2.5 Containarization

Most computer code is not self-contained but needs libraries and other software to work (e.g. the R programming language or packages). These external dependencies pose a risk for reproducibility because it may not be clear what besides the code and data is necessary and how to install it. Even when all needed software and their exact versions are recorded meticulously, it may be a challenge to install them. First, it is difficult to maintain different software versions on the same computer and second it may be unclear how to obtain an exact copy of some years old software version. Setting up a computer exactly as someone else is difficult enough, but replicating some

other computer how it was years ago is at best painstaking.

To overcome this challenge, the software environment of a project needs separation from the rest of the software environment. Technically such separation is called virtualization because one software environment is hosted on another. Such virtual environment allows each project to have its own software environment without interfering with each other. Hence, such setup is ideal for conservation and can be easily recreated on another machine.

Docker allows virtualization of the whole software stack down to the operating system, but in a much more lightweight way than traditional virtual machines. This lightweight but comprehensive virtualization is called containerization. Containers save storage by being based on each other enabling reuse. Hence if two containers are based, e.g. on a container for the same R version, they only use the storage they need for the different R packages. Containers are created from a plain specification called `Dockerfile`, that defines on which container it should be based and what software should be installed within it.

The `repro`-package automatically infers which packages are needed and creates an appropriate Dockerfiles and the container from it.

References

Andersen, H., & Hepburn, B. (2016). Scientific method. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2016). https://plato.stanford.edu/archives/sum2016/entries/scientific-method/; Metaphysics Research Lab, Stanford University.

Announcement: Reducing Our Irreproducibility. (2013). *Nature*, *496*(7446), 398–398. https://doi.org/10.1038/496398a

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., O'Neill, O., Rawlins, M.,

Thornton, J., & Vallance, P. (2012). Science as an open enterprise. *The Royal Society*.

Claerbout, J. F., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. *SEG Technical Program Expanded Abstracts 1992*, 601–604. https://doi.org/10.1190/1.1822162

Deutsche Forschungsgemeinschaft. (2019). *Leitlinien zur Sicherung guter wissenschaftlicher Praxis*. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf

Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145–155. https://doi.org/https://doi.org/10.1177/2515245919847421

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. In D. Kaplan, *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 392–409). SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n21

Gilbert, S. W. (1991). Model building and a definition of science. *Journal of Research in Science Teaching*, *28*(1), 73–79. https://doi.org/10.1002/tea.3660280107

Heilbron, J. L. (Ed.). (2004). The Oxford Companion to the History of Modern Science. *Reference Reviews*, *18*(4), 40–41. https://doi.org/10.1108/09504120410535443

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726. https://doi.org/10.1126/science.359.6377.725

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects.

Science, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kraker, P., Leony, D., Reinhardt, W., Gü, N., & Beham, nter. (2011). The case for an open science in technology enhanced learning. *International Journal of Technology Enhanced Learning*, *3*(6), 643. https://doi.org/10.1504/IJTEL.2011.045454

Martin, R. C. (2011). *The clean coder: A code of conduct for professional programmers / Robert C. Martin* (1. print.). Prentice Hall.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Müller, K. (2017). *Here: A simpler way to find your files*. https://CRAN.R-project.org/package=here

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Sci-*

ence, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Peikert, A. (2020). *Repro: Easy setup of a reproducible workflow*. https://github.com/aaronpeikert/repro

Peikert, A., & Brandmaier, A. M. (2019). *A Reproducible Data Analysis Workflow with R Markdown, Git, Make, and Docker* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/8xzqy

Popper, K. R. (1962). Some comments on truth and the growth of knowledge. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, Methodology and Philosophy of Science Proceedings of the 1960 International Congress* (Vol. 155). Stanford University Press.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1–12. https://doi.org/10.1037/a0018326

The Turing Way Community, Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., & Whitaker, K. (2019). *The Turing Way: A Handbook for Reproducible Data Science*. https://doi.org/10.5281/zenodo.3233986

Tichỳ, P. (1976). Verisimilitude redefined. *The British Journal for the Philosophy of Science*, *27*(1), 25–42.

Williams, J. M. (2017). *Style: Lessons in clarity and grace* (Twelfth Edition). Pearson.

Xie, Y., Allaire, J. J., & Grolemund, G. (2019). *R Markdown: The definitive guide*.