Reproducibility made simple

Automating reproducible research workflows

Aaron Peikert

2020-03-12

Contents

Abstract

This is a short summary.

1 Introduction

Claerbout & Karrenbach (1992) define reproducibility as the ability to gain the same results, from the same dataset. In contrast, 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 necessary. 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. Science should provide facts general enough to be theoretically verifiable by anyone is an argument deeply persuasive to me. Some have even argued that this 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 secondly, 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 given above. 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" here 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. Because 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.

This much more demanding standard of reproducibility may gain 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, @meehlTheoreticalRisksTabular1978). 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 call 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 facilitates replication the most 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.

Reproducibility is nothing special 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 (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 just 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 chunk of the researchers time, which they may better spend at advancing research. This additional effort offsets the inrease 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 & Brandmaier, 2020).

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

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

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

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., & Brandmaier, 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.

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

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