Statistical Science
2023, Vol. 38, No. 4, 525–526
https://doi.org/10.1214/23-STS909

Editorial: Special Issue on Reproducibility and Replicability

Alicia L. Carriquiry, Michael J. Daniels and Nancy Reid

There has been in the recent statistical literature a vigorous debate about the role of statistical methods in ensuring reproducibility and replicability of scientific studies. While this discussion has been part of our discipline for many decades, it has gained new urgency with the rapid increase in the size and scope of data relevant to nearly every discipline of academic study, as well as to government, non-governmental organizations, and industry. One aspect of this is the emphasis in the relatively new field of data science on the development of principles, strategies and software to enable reproducibility of published studies. Several statistical and scientific journals now insist on code and data being made available to reviewers, for example.

In this editorial we follow the National Academies' consensus study report [5] (NASEM, 2019) and use 'reproducibility' to mean obtaining consistent results using the same data and methods, and 'replicability' to mean obtaining consistent results across studies in similar or closely related settings.

What might be called the 'recent' literature on statistical aspects of reproducibility and replicability often takes as its starting point the highly cited article of Ioannidis (2005) [2], arguing that most published research results are neither reproducible nor replicable. Leek and Jager (2017) [4] study this quantitatively and come to a somewhat different conclusion. There have also been many calls for changes in statistical methods in order to enhance reproducibility or replicability, including suggestions to change the method for determining a declaration of statistical significance, to abandon declarations of statistical significance, to develop formal statistical guidelines for authors (Harrington et al., 2019 [1]; JASA, 2020 [3]), and more. In fall 2020, the Harvard Data Science Review published a special issue on reproducibility and replicability (Stodden, 2020 [6]), which included a summary of NASEM (2019) [5].

Alicia Carriquiry is Distinguished Professor and President's Chair and Director of CSAFE, Department of Statistics, Iowa State University, Ames, Iowa 50011, USA (e-mail: alicia@iastate.edu). Mike Daniels is Professor and Chair, Andrew Banks Family Endowed Chair, Department of Statistics, University of Florida, Gainesville, Florida 32603, USA (e-mail: daniels@ufl.edu). Nancy Reid is University Professor, Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5G 1X6, Canada (e-mail: nancym.reid@utoronto.ca).

Taking a very broad view, issues of reproducibility and replicability touch on almost all areas of science, and can only be addressed through concerted efforts at an institutional level. Taking a very narrow view of just the statistical aspects of reproducibility and replicability still leaves a great deal of scope for discussion and disagreement. Recognizing the importance of the debates, the editor of *Statistical Science* asked us to consider what our flagship review journal could contribute to the discussion. We strove to address the issues through a mix of papers on theory and on applications, highlighting particular application areas where the problems struck us as especially interesting to our readership, and focussing the theoretical work on multiple testing, post-selection inference, and methods that provide statistical guarantees without detailed model assumptions.

The papers in this volume cover various aspects of both theory and application. They are generally concerned with replicability, which is arguably more relevant for study of the theory and methods of statistical science. It should be noted that several authors did indeed provide code and data to ensure their computational results are reproducible.

Rothenhausler and Bülhmann discuss a general approach to both stability and generalizability of inferences. From a theoretical point of view, internal stability to perturbations of the data distribution helps to ensure replicability of findings. External validity is discussed in the context of both point estimation and uncertainty quantification. Parmigiani explores replicability for predictions, an important concern in machine learning. He characterizes this as results that are consistent across studies suitable to address the same prediction question. He proposes a multi-agent framework for defining replicability and shows that some of the common practical approaches are special cases.

Robertson, Wason and Ramdas focus their attention on large-scale hypothesis testing in online settings, which gives rise to issues of multiplicity, and thus affects replicability if these issues are not addressed. Examples treated include A/B testing, platform trials in which several treatments use the same control group, and the use by many groups of researchers of the same online database. They describe and illustrate several algorithms for control of error rates that explicitly depend on time, such as the family-wise error rate $FWER(t)$ and the false-discovery

rate $FDR(t)$. Ramdas, Grünwald, Vovk and Shafer highlight recent work on methods of inference that have universal guarantees, irrespective of the model. They emphasize the links to betting and game theory, and describe how their theory of E-values can be used at arbitrary stopping times. They discuss a recent theoretical advances related to universal inference, that extend the application of E-values and E-processes to complex settings.

Bogomolov and Heller discuss the problem of findings from meta-analysis being completely driven by a single study and thus being non-replicable. They provide an overview of analyses, with the appropriate theory, that can be used to establish replicability in the context of a single outcome in multiple studies and multiple outcomes in multiple studies. Freuli, Held and Heyard present a detailed simulation study to consider various metrics of replication success, in the presence of what they call "questionable research practices". The metrics for success include conventional guides, such as the two-trials rule and meta-analytic approaches, two metrics based on the sceptical p-value, and a replication metric based on a Bayes factor. The questionable research practices include interim and subgroup analyses, selection of significant results, and selection of covariates.

Branter, Chang, Nguyen, Hong, Di Stefano and Stuart provide a comprehensive review of methods for integrating randomized trials and observational studies in three different data scenarios: aggregate-level data, federated learning, and individual participant-level data. They emphasize the importance of understanding how the original data were collected, analyzed, and presented, to help ensure replicability of the treatment effect heterogeneity findings.

Possolo focuses on the important topic of measurement, and the often-overlooked impact of inter-laboratory heterogeneity on reproducibility of studies. Through a collection of examples, Possolo discusses some of the statistical challenges that arise when comparing and synthesizing results obtained by individual investigators, and highlights situations where the same data, subject to slightly different modeling assumptions, lead to substantively different conclusions.

There is a great deal of outstanding research relevant to reproducibility and replicability that we have omitted: we could provide just a snapshot of the field in this special issue. We tried to use a wide, albeit subjective, lens in the hopes of providing a broad overview of a very important set of problems.

REFERENCES

[1] HARRINGTON, D., D'AGOSTINO, R. B., GATSONIS, C., HOGAN, J. W., HUNTER, D. J., NORMAND, S.-L. T., DRAZEN, J. M. and HAMEL, M. B. (2019). New guidelines for statistical reporting in the. *N. Engl. J. Med.* **381** 285–286. https://doi.org/10.1056/NEJMe1906559

[2] IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *PLoS Med* **2** e124. https://doi.org/10.1371/journal.pmed.0020124

[3] JASA Reproducibility Guide (2020). https://jasa-acs.github.io/repro-guide/.

[4] LEEK, T. J. and JAGER, L. R. (2017). Is most published research really false? *Annu. Rev. Stat. Appl.* **4** 109–122. https://doi.org/10.1146/annurev-statistics-060116-054104

[5] National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303

[6] STODDEN, V. (2020). Theme editor's introduction to reproducibility and replicability in science. *Harv. Data Sci. Rev.* **2** 4.