Reproducibility and Error

Charles J. Geyer

School of Statistics and Minnesota Center for Philosophy of Science,
University of Minnesota

May 8, 2025

Abstract

Reproducible research and scientific error are viewed through the lens of the personal experience of one statistician having seventeen years of experience following the precepts of reproducible research. The shocking assertion will be that most statistics in most scientific papers has errors. Few authors, referees, editors, or readers of scientific papers pay adequate attention to statistics. Statistics done is often so poorly described that readers can have no confidence it was done validly. We also discuss other statistical reasons for the replication crisis: multiple testing without correction and publication bias. And we discuss best practices: complete reproducibility of all computations starting from raw data, version control, code reviews, permanent repositories, software packages, software testing, literate programming, and data cleaning.

1 Introduction

It is now commonplace that science has a "replication crisis" or "reproducibility crisis" (Google has a hundred times more hits for the former phrase than the latter). There have been many attempts to distinguish replication from reproduction and distinguish subtypes of each. This paper will not try anything of the sort.

Your humble author is a statistician. This paper lays out one statistician's take on reproducibility. I have been doing, teaching, and mentoring reproducible research since 2005. I keep learning. Some of what I say here was learned while this was being written. Some was learned long before reproducibility achieved widespread attention.

We limit ourselves to aspects of reproducibility about which statisticians have something to say. Statisticians tend not to know their limitations. R. A. Fisher wrote a book *Design of Experiments* (Fisher, 1935, ninth and last edition 1971) that was so important that many statistics departments name a course like that, and from that one might infer that statisticians think they know everything about scientific experiments. I hereby explicitly disclaim that. This paper is about some things involved in the reproducibility crisis, not all things. There are many aspects of good scientific practice that we do not touch.

Here are the ideas of this paper.

1. Most scientific papers that need statistics have conclusions that are not actually supported by the statistical calculations done, because of
   (a) mathematical or computational error,
   (b) statistical procedures inappropriate for the data, or
   (c) statistical procedures that do not lead to the inferences claimed.
2. Good computing practices — version control, well thought out testing, code reviews, literate programming — are essential to correct computing.
3. Failure to do all calculations from raw data to conclusions (every number or figure shown in a paper) in a way that is fully reproducible and available in a permanent public repository is, by itself, a questionable research practice.
4. Failure to do statistics as if it could have been pre-registered is a questionable research practice.
5. Journals that use $P < 0.05$ as a criterion of publication are not scientific journals (publishing only one side of a story is as unscientific as it is possible to be).
6. Statistics should be adequately described, at least in the supplementary material.
7. Scientific papers whose conclusions depend on nontrivial statistics should have statistical referees, and those referees should be heeded.
8. Not all errors are describable by statistics. There is also what physicists call *systematic error* that is the same in every replication of an

experiment. Physicists regularly attempt to quantify this. Others should too.

Points 1 and 2 are the most shocking. They are certainly not the conventional wisdom, and we cannot prove them. But the "reproducibility crisis" had no proof until recently because no one was trying to prove it. Nothing in the "reproducibility crisis" was a surprise. Its every cause is scientists ignoring how experts have always said that science should be done. What points 1 and 2 say is that scientists and their journals are still ignoring how experts say statistics and computing should be done. Thus, in all probability, if anyone looked hard, they would find the literature riddled with statistical and computing errors.

Points 3 and 4 are not controversial, being now widely accepted in principle, but still not widely accepted in practice. Point 5 is still somewhat controversial: most journals do not yet say this.

Point 3 is especially important. The old-fashioned way of dealing with computation — readers contact authors to ask for computer code and maybe authors respond and maybe they send something useful — is no longer acceptable. Science is supposed to be reproducible — all of it. Many journals now require that data be put in a permanent public repository. Few journals require that all computer code be treated similarly, but all should.

Points 4 and 5 are about multiple testing without correction and publication bias, which are well understood to be problematic and major causes of the reproducibility crisis. Your humble author gave a talk in 2012 before reproducibility was called a crisis that said that these were the main causes of what would later be called the reproducibility crisis (Geyer, 2012).

That talk said that pre-registration (putting the protocol for a study in a permanent repository before the study is done, `https://www.cos.io/initiatives/prereg`) was a solution to multiple testing without correction and all problems that statisticians call "data snooping" or "data dredging" and I call playing statistics like playing tennis without a net. That talk did not envisage that pre-registration would become as widespread as it has. That talk did not guess that pre-acceptance (accepting a paper before the experiment was done) would be a complete solution to the publication bias problem (for papers under this system).

The statistical analysis for every paper should be done as if it could have been pre-registered — the analysis chosen before the data were collected. And every paper should be refereed as if it were considered for pre-acceptance — the decision should only be about whether the experiment or observational study was done well — *not whether the results were*

positive or negative.

Points 6 and 7 refer to a bizarre trait of scientists: most do not consider statistics a part of the science. A scientific paper should completely describe the materials and methods, in the supplementary material if the description is too long for the paper itself, but most authors do not include the statistics here. They have almost casual regard for statistics, paying no attention to details as if they don't matter. This leads to the errors described in point 1. Hence our recommendations in points 6 and 7.

Point 8 is about what statistics cannot do: deal with nonrandom error.

2 Reproducible Computing

2.1 Then

Your humble author has not always been a convert to point 3. The first paper drafted by your humble author with fully reproducible computing is Geyer, Wagenius, and Shaw (2007, first submitted 2005). The computer code for this paper is mostly in R package `aster` (Geyer, 2023, version 0.2 appeared on CRAN 2005-07-20) which is a package for the statistical computer language R available from the public repository CRAN (`https://cran.r-project.org`).

When one tries to install a package from CRAN, one gets the current version, but all previous versions are available for download and installation by those who know how to install from source (documented in the book *R Installation and Administration* installed with every installation of R; type `help.search()` in R and click on the link for this book). Even packages that have been removed from CRAN have archived versions available. Source code for all versions is available for inspection. Thus CRAN satisfies the requirements to be a *permanent* public repository.

Since the R language and the `aster` package are free software, any user can install the software on any computer in minutes and use it freely for any purpose. If one does not like R, then there are similar public software repositories for other computer languages. R is just the computer language of choice for statisticians.

Of course, R package `aster` only contains computer code useful for all users. It does not contain all of the code used for the statistical analysis in Geyer, et al. (2007). That was put in a technical report (Geyer, et al., 2005). In a failure of perfect reproducibility, that technical report was not placed in a permanent public repository until 2018. It was on the internet at `https://www.stat.umn.edu/geyer/aster/tr644.pdf` but that file was

controlled by your humble author, so readers have to trust that this was not edited after its publication date (the journal did not require a permanent repository at the time, nor did I understand the importance of this at the time).

2.2 Now

Now we know how to do this better. The paper Kulbaba, et al. (2019) needed a correction, and this correction (Geyer, et al., 2022) was done fully reproducibly. The original paper was done almost but not quite fully reproducibly. All of the code for the calculations was placed in a public repository Kulbaba (2019) but that repository was not *permanent* because it was editable by the owner Kulbaba. For the correction, the story was similar but different. All of the code for the calculations was placed under version control in a GitHub repository Geyer (2022), and that is (like the GitHub repository for the original paper) not permanent because editable by the owner Geyer. But for this correction we took the additional step of placing snapshots of the repository on Zenodo (`https://zenodo.org/`), which is a permanent repository. Zenodo does allow updates and corrections. But all previous versions are shown and are permanently available. (In fact, we had to make three versions, only the last being correct.) So if, in the future, further corrections are needed, they can be done. Zenodo gives not only permanence but also a DOI (document object identifier) that always resolves to the most recent version (and also links to earlier versions). Once one has "enabled" a GitHub repository for Zenodo, every time a "release" is made on GitHub, this becomes a permanent version on Zenodo.

We now think this is the best way to do reproducible computing. We can use literate programming (R markdown). We can use version control (`git`). We can make our code public (GitHub). We can also make certain commits permanent (Zenodo). And we get a DOI to cite.

It is interesting that this correction exists because the original paper was done reproducibly. The story is told in the acknowledgments in the correction document. A researcher (Anna Peschel) while trying to follow the archived computer code for the original paper noticed "that estimates were not corrected for subsampling" and reported the problem. The correction resolves the issue. If the original paper had not been done reproducibly, there would have been no evidence of an error to correct.

2.3 Literate Programming

Literate programming (Knuth, 1984) was a new idea in computing following decades of computer code being "documented" by comments in the code and completely separate articles and books. The problems with the previous system were widely recognized. The comments in the code were unreadable and often incorrect (because programmers paid so little attention to them). The code cut and pasted into other documents was often incorrect or even just broken (because it was never actually executed).

The literate programming solution was to merge documents and code so the code in the document is *actually executed* to *actually produce* the output shown in the document (whether numeric results, tables, or figures).

In our opinion, this is the only valid way to present computer code understandably and reproducibly. The R statistical computing language that we use for our work has four literate programming systems: `Sweave` (Leisch, 2002), `knitr` (Xie, 2015, 2022, first appeared 2012), R markdown (Xie, 2019; Allaire, et al., 2022, first appeared 2014), and Quarto (`https://quarto.org/`, first appeared 2021).

`Sweave` is more or less obsolete, replaced by `knitr`. Both are considered somewhat hard to use because they require knowledge of LaTeX. R markdown is easier to use. Over time I have gone from `Sweave` to `knitr` to R markdown. Even though R markdown gives you much less control over how your document looks, it provides a much easier to understand example for others. And one of the main reasons I make R markdown documents is to provide an example for students and scientists (for more see, `https://rmarkdown.rstudio.com/`). Quarto is R markdown with additional features.

We wouldn't think that a scientific paper that made claims with no explanation or justification would be worth reading. The same goes for computer code. We need to see the code, that it does what it is claimed to do, and the explanation and justification.

2.4 Summary

In between the two papers described above (one a correction) your humble author has made every paper for which he was the lead author or lead data analyst or supervisor of the lead author or lead data analyst fully reproducible (17 years of experience with reproducibility).

We have found that reproducibility is not only valuable *per se* but also serves as a resource for teaching researchers how to use the methods. One

might think that no one reads these supplementary documents and the computer code in them, but one would be wrong. They serve as examples for many researchers. If you want to use the methods of the papers, that's where you find out how to.

3 Correct Computing

Scientific computing is odd. Some of it is very professionally done. Think of data collection and storage for the Large Hadron Collider (https://home.cern/science/computing). But most of scientific computing is done by scientists with no formal training in computing and often with very little experience.

Professional computing, also called software engineering, has trouble. The term "software crisis" was coined in 1968. It is still ongoing (Fitzgerald, 2012). For all of the advances in programming practices — version control, formal quality control, code reviews, much improved and safer programming languages — programming is still very hard (Brooks, 1987). There are no tricks that can make the hard problems, due to essential complexity, go away. Even the best computing companies, employing thousands of the best programmers available, produce software that is buggy, is expensive, is hard to use, does not do what is expected, and is delivered late or never.

If scientists who are amateur programmers do not use any of the tools and methods required to get somewhat acceptable software from professional programmers, what reason is there to believe that anything they do is correct — especially if they give no evidence? I claim there is no reason to trust the computer code underlying most scientific research — even when that computing is done completely reproducibly. You can trust it, after you have checked it.

Reproducible computing is some weak evidence. At least it makes the code possible to check. But referees of scientific papers almost never check code. So the fact that a paper has appeared in the refereed literature gives one zero reason to trust any computer code that underlies the paper. You have to check it yourself.

Use of good software engineering practices is stronger evidence of good computing. Scientists doing computing should use safe computer languages, literate programming, version control, quality control, and code reviews. These sorts of things never appear in the "materials and methods" sections of scientific papers or even in the "supplementary material." They should.

I cannot say that I use everything known from software engineering. I

have used version control since 2005. R packages can have formal tests run automatically whenever the package is checked. Mine do. The tests should check that the software works correctly on some rather generic problems. Mine do. I generally do not have independent quality control by other people. But it would be good if I did. I have done some code reviews with some students. They work very well to find parts of the code that are not well justified and may contain errors. Often they do find errors.

Software testing is hard. But one thing should be obvious. If no testing is done, there is no reason to believe the software is correct. People do make mistakes, and scientists are people too. Not looking for mistakes means that mistakes are there undiscovered. There is never a reason to assume that people are perfect.

4 Statistics

4.1 Scientific Culture

It seems bizarre to a statistician that most scientists are very casual about statistics. I have a joke that to most scientists $P < 0.05$ means "statistics has proved that every idea I have ever had on this subject is correct." No scientist would be brazen enough to say something like that, but many seem to act as if they think that. No details of the statistics seem to have any importance to their thinking.

Many statistical hypothesis tests have null hypotheses that are mere straw men to knock down. When the test duly knocks them down, the proper interpretation is that the test has shown that *something* is going on in the data, but not necessarily that the scientific theory is true (except in the simplest possible case when the test is about treatment effect or no treatment effect, then the test can show a treatment effect is statistically significant, although not necessarily large enough to be scientifically significant). But "these data are not worthless" is not a strong story, so scientists routinely go beyond what the statistics actually says.

Here's a story. Cowley and Atchley (1992) defend their use of matrix permutation tests against the criticism of Shaw (1992) that the mathematical assumptions required for a permutation test — that the random variables being permuted have an exchangeable joint distribution — is simply ludicrous in the application being discussed. Cowley and Atchley are permuting rows and columns of estimated genetic variance covariance matrices. These rows and columns correspond to phenotypic traits of some organism, say height, weight, and G6PD activity (G6PD is an enzyme). The mathemat-

ics requires that, if you permute the variables, that does not change the probability distribution. Height and weight have *exactly* the same distribution. Height and G6PD activity have *exactly* the same distribution. Height and weight have *exactly* the same correlation as weight and G6PD activity. When you reject this null hypothesis, you have not learned anything. Height and weight are different? Who knew? You can publish things in a first-tier scientific journal that are as ridiculous as $2 + 2 = 5$. To many scientists, some of whom are editors and referees, mathematical correctness is not an issue.

Here are some more stories that lack citations because the dispute never made it into print for reasons that show various pathologies in the culture of academic science.

An associate editor (AE) of a scientific journal, who was a statistician, recommended that a paper be rejected because the statistical analysis was inappropriate for the scientific issues under discussion. The editor initially agreed. The authors protested that they used SAS (https://www.sas.com/). The editor reversed the decision, and the paper was published. The AE resigned; no point in having anything more to do with that journal.

An AE of a first tier scientific journal, who was a biologist, sent a paper that involved a lot of statistics to me for refereeing. The mathematics was nonsense. It required assuming that a lot of variables were stochastically independent that were reported to be very highly correlated. Again this is a wrong as $2+2 = 5$. The AE recommended the paper be rejected. The editor initially agreed. One of the authors protested that he was an international big cheese, had published multiple papers in the area, and they couldn't do this to him (no argument about the math). The editor reversed the decision, and the paper was published. The AE protested at a meeting of the society that runs the journal, but lost. As my department chair said, "if societies second guessed editors, they wouldn't have editors."

Once these decisions were made there is not a lot one can do. Picking a public fight with a first-tier journal and perhaps with the society owning it is not going to make one a lot of friends, especially when the argument is a technical one that many scientists will not understand.

This goes along with a very strong bias towards positive results. This will be discussed under the notion of publication bias below, but it goes far beyond that. Most journals will not publish a paper that says nothing positive, just that another paper is wrong. To actually publish on these subjects, we would have had to derail our research and figure out how to do what those authors were trying to do *correctly* and without their assistance, a thankless and perhaps impossible task. Better to recognize a no win

situation.

Of course, most scientific papers do not come with stories like this. But it is very common for statistics done to be so sketchily described that no one could have any idea whether it is valid or appropriate. Hence our points 6 and 7.

And finally, a quote from a letter from a referee responding to authors' reply to the referees initial comments about suspect statistics.

> I understand that, as it is the case for many scientists, statistics is (too often) considered just a somewhat auxiliary aspect and maybe far from the main point of a manuscript. But allow me to point out that, if the statistical analysis is not done correctly, even what may appear to be the most groundbreaking result, effectively reduces to just a falsity. The intent of my comments is to give the authors the chance to prove that they are in the former category and not in the latter.

Scientists should not have to be told this, but they do.

4.2 Data Dredging

Many scientists have wondered around in data searching for $P < 0.05$. This is something statisticians have long disparaged as "data snooping" or "data dredging". There are areas of statistics called multiple comparisons (Hsu, 1996), model comparison (Burnham and Anderson, 2002), and model averaging (Hoeting et al., 1999; Claeskens and Hjort, 2008) that replace ad hoc data dredging with valid statistics. But they take all the fun out of data dredging. If you don't use these valid methods, you get

> Remember Barnet Woolf's definition of statistics as that branch of mathematics which enables a man to do twenty experiments a year and publish one false result in *Nature* (Maynard Smith, 1986).

Putting new wine in old bottles, data dredging has recently been called P-hacking and HARKing, but these focus attention on the wrong issues. Cherry-picking evidence of all kinds is bad science, not only cherry picking P-values or hypotheses (more on this in Section 4.3 below).

Pre-registration ties the hands of authors of scientific papers. They must do the pre-registered statistical analysis and report it (otherwise the referees should object), and the pre-registration should have been done before

the data were collected. This means no cherry-picking is possible. Pre-registration is still not the norm. This means that reproducible statistical analysis must accurately report all cherry-picking that was done. If a lot is reported, knowledgeable readers are given enough information to see exactly how bogus this is and dismiss the claims of the paper. Naive readers are fooled. I call this "honest cheating" — honest because it is accurately and completely described, cheating because it is playing statistics like playing tennis without a net.

In many areas of science this "honest cheating" is still the norm. The 2012 talk (Geyer, 2012) uses as an example the kerfuffle about electric power lines putatively causing cancer, which, as far as I can see, was entirely due to multiple testing without correction. Every paper that was bad science tested for association with more than 20 different cancers and found at least one, of course, not always the same one. Every paper that eschewed multiple testing without correction found none. Yet the National Research Council report on the subject just barely mentioned multiple testing without correction in its 379 page report on the subject. Most scientists are uncomfortable saying that bad statistics ruins science. They don't want to make statistics that important. They may have some bad statistics in their own research.

Data dredging is just as bad when it is done by a whole academic discipline rather than authors of single papers. This is called publication bias. When referees and editors only accept papers that say $P < 0.05$ or only have "positive results" by some indication, this causes a very serious problem in the literature. Telling only one side of the story is as anti-scientific as one can get. Of course, referees and editors don't *think* they are being anti-scientific. But they are.

Pre-acceptance ties the hands of editors and referees of scientific papers. If the design of the experiment or observational study is good, then the paper is tentatively accepted before any data are collected. So long as the authors do the experiment and report as they said they would, the paper is automatically accepted *regardless of whether the results are positive or negative*. (And even the notion that scientific results can be positive or negative is weird. Positive means showing that some hypothesis is correct — hypothesis-driven research strikes again.)

Pre-registration and pre-acceptance are still not the norm. But all papers should be written, refereed, and edited as if they could have been. Authors should not data dredge. Referees and editors should not judge papers on whether the results are positive or negative.

4.3 Hypothesis Testing Obsolete?

Many scientists who do not understand statistics have called for the abandonment of statistical hypothesis tests and P-values in favor of confidence intervals or Bayesian inference (Berner and Amrhein, 2022, and references cited therein). Statisticians do not agree (Wasserstein and Lazar, 2016).

In fact, hypothesis tests and confidence intervals are the same mathematics looked at in two different ways. It is nonsense to say that one is valid and not the other. Multiple testing without correction is playing statistics like playing tennis without a net. So is multiple confidence intervals without correction (same math). Any cherry picking of evidence is bad. Cherry picking ruins Bayesian statistics in the same way it ruins frequentist statistics (Rosenbaum and Rubin, 1984; de Heide and Grünwald, 2020). Cherry picking is bad even if it does not involve statistics at all (Feynman, 1985, Chapter "Cargo Cult Science").

Thinking that hypothesis tests or P-values are the difficulty, rather than cherry picking of evidence, completely mischaracterizes the problem.

4.4 Assumptions and Interpretations

Statistical procedures have assumptions. When those assumptions are grossly violated, the statistics is nonsense. Statistical procedures have interpretations. When those interpretations are ignored, the statistics often does not actually support the scientific conclusions it is claimed to support. This is point of the stories in Section 4.1 above.

4.5 The Crowdsourced Replication Initiative

As this was being written, some evidence appeared of the variability of statistical analysis done by scientists (Crowdsourced Replication Initiative, 2022). 73 research teams analyzed the same social science data investigating the same research hypothesis. Together they fit 1261 models and came to conclusions that varied widely. The meta-analysis done by the principal investigators could not account for the variability of the results. They say it is not due to data dredging, although it is unclear to me how reliable this conclusion is. The meta-analysis was done reproducibly, but (as far as I can see), the initial analyses by the 73 research teams was not done reproducibly and cannot be checked for errors. The meta-analysis did not investigate how much of the variability of results was due to outright error, They do claim

that the 73 research teams were not naive; many had taught courses in the type of analysis they were doing.

In conclusion, there is no conclusion. This should be very worrying, but no explanation of these results will be forthcoming. We need a similar project in which all of the analyses were done completely reproducibly. Then, with enough work, it could be discovered what accounts for the variability of results.

The study referenced above is only one of many. For another, see Kummerfeld and Jones (2023, and this volume).

5 Systematic Error

Experimental physicists routinely consider not only statistical error (variability of repeated measurements) but also *systematic error* (measurements not correctly measuring the thing being measured). Statistics is no help with systematic error. Every measurement is biased in the same way.

The only way to get a handle on systematic error is to use knowledge of theory and practice to bound how wrong measurements can be. Of course, if what one is measuring is defined purely operationally (IQ is what IQ tests measure), then there can be no systematic error (but one has a different notion of IQ for each different test).

In physics we do have theoretical definitions of most concepts, and measurements can thus have systematic error. We also have theoretical definitions in other sciences. They could pay more attention to systematic error than they do.

References

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2022). R package `rmarkdown`: Dynamic Documents for R, version 2.16. https://rmarkdown.rstudio.com. https://cran.r-project.org/package=rmarkdown.

Berner, D., and Amrhein, V. (2022). Why and how we should join the shift from significance testing to estimation. *Journal of Evolutionary Biology*, **35**, 777–787. doi:https://doi.org/10.1111/jeb.14009.

Brooks, Jr., F. P. (1987). No silver bullet: Essence and accidents of software engineering. *Computer*, **20**, 10–19.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition. Springer, New York.

Claeskens, G., and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Cowley, D. E., and Atchley, W. R. (1992). Comparison of quantitative genetic parameters. *Evolution*, **46**, 1965–1967. doi:10.1111/j.1558-5646.1992.tb01184.x.

Crowdsourced Replication Initiative (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. To appear in *Proceedings of the National Academy of Sciences of the United States of America*. `https://github.com/nbreznau/CRI`. There are 166 authors. The research design and analysis was by Nate Breznau, Eike Mark Rinke, Alexander Wuttke, and Hung H.V. Nguyen. Other authors were researchers whose analyses went into the meta-analysis.

de Heide, R., and Grünwald, P. D. (2020). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, **28**, 795–812. doi:10.3758/s13423-020-01803-x.

Feynman, R. P. (1985). *Surely You're Joking, Mr. Feynman!*. Bantam Books, New York.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyde, Edinburgh.

Fitzgerald, B. (2012). Software Crisis 2.0. *Computer*, **45**, 89–91. doi:10.1109/MC.2012.147.

Geyer, C. J. (2012). Slides for the SST (Science and Technology Studies) Annual Science Studies Symposium, University of Minnesota. `http://users.stat.umn.edu/~geyer/101b.pdf`.

Geyer, C. J. (2023). R package `aster`: Aster Models, version 1.1-3. `https://cran.r-project.org/package=aster`.

Geyer, C. J. (2022). GitHub package `Evolution-correction`. `https://github.com/cjgeyer/Evolution-correction`

Geyer, C. J., Kulbaba, M. W., Sheth, S. N., Pain, R. E., Eckhart, V. M., and Shaw, R. G. (2022). Correction for Kulbaba et al. (2019). *Evolution*,

76, 3074. doi:10.1111/evo.14607. Supplementary material, version 2.0.1. doi:10.5281/zenodo.7013098.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2005). Aster Models for Life History Analysis. Technical Report No. 644. School of Statistics, University of Minnesota. https://hdl.handle.net/11299/199666.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, **94**, 415–426. doi:10.1093/biomet/asm030.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, **19**, 382–417. doi:10.1214/ss/1009212519. The printed version has numerous typos that were the fault of the journal and are corrected in http://www.stat.washington.edu/www/research/online/1999/hoeting.pdf.

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC, Boca Raton.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, **2**, e124. doi:10.1371/journal.pmed.0020124.

Knuth, D. E. (1984). Literate programming. *Computer Journal*, **27**, 97–111. doi:10.1093/comjnl/27.2.97.

Kulbaba, M. W. (2019). GitHub repository `adaptive-capacity`. https://github.com/mason-kulbaba/adaptive-capacity.

Kulbaba, M. W., Sheth, S. N., Pain, R. E., Eckhart, V. M., and Shaw, R. G. (2019). Additive genetic variance for lifetime fitness and the capacity for adaptation in an annual plant. *Evolution*, **73**, 1746–1758. doi:10.1111/evo.13830.

Kummerfeld, E., and Jones, G. L. (2023). One data set, many analysts: Implications for practicing scientists. *Frontiers in Psychology*, **14**, 6 pages. doi:10.3389/fpsyg.2023.1094150.

Leisch, F. (2002). Sweave, part I: Mixing R and LaTeX. *R News*, **2**, 28–31. https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf#page=28.

Maynard Smith, J. (1986). Molecules are not enough. *London Review of Books*, **8**, number 2.

Rosenbaum, P. R., and Rubin, D. B. (1984). Sensitivity of Bayes inference with data-dependent stopping rules. *American Statistician*, **38**, 106–109. doi:10.1080/00031305.1984.10483176.

Shaw, R. G. (1992). Comparison of quantitative genetic parameters: Reply to Cowley and Atchley. *Evolution*, **46**, 1967–1969. doi:10.1111/j.1558-5646.1992.tb01185.x.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *American Statistician*, **70**, 129–133. doi:10.1080/00031305.2016.1154108.

Wikipedia contributors (2022), Literate programming. In *Wikipedia, The Free Encyclopedia*. Retrieved 14:59, September 24, 2022. https://en.wikipedia.org/w/index.php?title=Literate_programming&oldid=1111131296.

Xie, Y. (2015). *Dynamic Documents with R and knitr*, 2nd edition. Chapman & Hall/CRC, Boca Raton.

Xie, Y. (2022). R package `knitr`: A General-Purpose Package for Dynamic Report Generation in R, version 1.40. https://cran.r-project.org/package=knitr.

Xie, Y., Allaire, J. J., and Grolemund, G. (2019). *R Markdown: The Definitive Guide*. Chapman & Hall/CRC, Boca Raton. https://bookdown.org/yihui/rmarkdown/.