

Cargo-cult statistics and scientific crisis



The mechanical, ritualistic application of statistics is contributing to a crisis in science. Education, software and peer review have encouraged poor practice – and it is time for statisticians to fight back. By **Philip B. Stark** and **Andrea Saltelli**

Poor practice is catching up with science, manifesting in part in the failure of results to be reproducible and replicable. Various causes have been posited, but we believe that poor statistical education and practice are symptoms of and contributors to problems in science as a whole.

The problem is one of *cargo-cult statistics* – the ritualistic miming of statistics rather than conscientious practice. This has become the norm in many disciplines, reinforced and abetted by statistical education, statistical software, and editorial policies.

At the risk of oversimplifying a complex historical process, we think the strongest force pushing science (and statistics) in the wrong direction is existential: science has become a career, rather than a calling, while quality control mechanisms have not kept pace.

Some, such as historian and sociologist of science Steven Shapin, still argue that science survives thanks to the ethical commitment of scientists,¹ but others, such as philosopher of science Jerome Ravetz, find this a charitable perspective.² Much of what is currently called “science” may be viewed as mechanical application of particular technologies, including statistical calculations, rather than adherence to shared moral norms.

We believe the root of the problem lies in the mid-twentieth century.

The bigger picture

After World War II, governments increased funding for science in response to the assessment that scientific progress is important for national security, prosperity, and quality of life. This increased the scale of science and the pool of scientific labour: science became “big



Philip B. Stark is professor in the Department of Statistics and associate dean in the Division of Mathematical and Physical Sciences at the University of California, Berkeley.



Andrea Saltelli is an adjunct professor in the Centre for the Study of the Sciences and the Humanities at the University of Bergen, and a researcher with the Open Evidence Research group at the Universitat Oberta de Catalunya, Barcelona.

science” conducted by career professionals. The resulting increase in scientific output required new approaches to managing science and scientists, including new government agencies and a focus on quantifying scientific productivity. The norms and self-regulating aspects of “little science” – communities that valued questioning, craftsmanship, scepticism, self-doubt, critical appraisal of the quality of evidence, and the verifiable, and verifiably replicable, advancement of human knowledge – gave way to current approaches centring on metrics, funding, publication, and prestige. Such approaches may invite and reward “gaming” of the system.

When understanding, care, and honesty become valued less than novelty, visibility, scale, funding, and salary, science is at risk. Elements of the present crisis were anticipated by scholars such as Price³ and Ravetz⁴ in the 1960s and 1970s; a more modern explanation of science’s crisis, in terms of the prevailing economic model – and of the commodification of science – is offered by Mirowski.⁵ Other scholars such as John Ioannidis, Brian Nosek, Marc Edwards, *et al.*, now study the perverse system of incentives and its consequences and how bad science outcompetes better science.^{6–8}

While some argue that there is no crisis (or at least no systemic problem), bad incentives, bad scientific practices, outdated methods of vetting and disseminating results, and technoscience appear to be producing misleading and incorrect results. This might produce a crisis of biblical proportions. As Edwards and Roy write: “If a critical mass of scientists become untrustworthy, a tipping point is possible in which the scientific enterprise itself becomes inherently corrupt and public trust is lost, risking a new dark age with devastating consequences to humanity.”⁹

Scientists collectively risk losing credibility and authority in part because of prominent examples of poor practice, but also because many are guilty of ultrarecrepitation: acting as if their stature in one domain makes them authoritative in others. Science is “show me”, not “trust me”. The example of 107 Nobel laureates – mostly in areas unrelated to genetics, agriculture, ecology, or public health – endorsing one side of the genetically modified organisms in food argument as “scientific” is a visible example of prestige and uninformed consensus conflated with evidence. As G. K. Chesterton wrote: “Fallacies do not cease to be fallacies because they become fashions.”¹⁰

Statistics was developed to root out error, appraise evidence, quantify uncertainty, and generally to keep us from fooling ourselves. However, increasingly often, it is used instead to aid and abet weak science – a role it can perform well when used mechanically or ritually.

of methods without understanding their assumptions, limitations, or interpretation will surely reduce scientific replicability. There are, of course, concerns about statistical practice. For instance, a statement on *p*-values by the American Statistical Association (ASA) was accompanied by no fewer than 21 commentaries, mostly by practitioners involved in drafting the ASA statement.¹² Their disagreement could be misinterpreted to suggest that anything goes in statistics, but diversity of opinion within statistics is not as broad as it may appear to outsiders.

The largest divide is between frequentist and Bayesian philosophies, which differ fundamentally in how they conceive of and quantify uncertainty, and even in their ontologies. But no good statistician, Bayesian or frequentist, would ignore how the data were generated in assessing statistical evidence nor would claim that a *p*-value is the probability that the null hypothesis is true – two common abuses.

The misuse of *p*-values, hypothesis tests, and confidence intervals might be deemed *frequentist* cargo-cult statistics. There is also *Bayesian* cargo-cult statistics. While a great deal of thought has been given to methods for eliciting priors, in practice, priors are often chosen for convenience or out of habit; perhaps worse, some practitioners choose the prior after looking at the data, trying several priors, and looking at the results – in which case Bayes’ rule no longer applies. Such practices make Bayesian data analysis a rote, conventional calculation rather than a circumspect application of probability theory and Bayesian philosophy.

Some scientists use an incoherent hotchpotch of Bayesian methods and frequentist measures of uncertainty in the same analysis, with no apparent understanding of the fundamental mathematical and philosophical incommensurability of Bayesian and frequentist measures of uncertainty. Some add the lengths of confidence intervals to the lengths of credible intervals or add systematic uncertainties in quadrature, as if they were independent random errors. Some conflate confidence levels with credible levels. We have seen examples in a number of disciplines, notably high-energy particle physics and cosmology.

To paraphrase David Freedman, much frequentist statistics is about what you would do if you had a model, and much Bayesian statistics is about what you would do if you

Many applications of statistics are cargo-cult statistics: practitioners go through the motions with scant understanding

Cargo-cult statistics

In his 1974 Caltech commencement speech, Nobel physicist Richard Feynman coined the label “cargo-cult science” for work that has some formal trappings of science but does not practise the scientific method.¹¹

Feynman’s neologism borrows from anthropological observations of Melanesian cultures that experienced a bonanza in World War II, when military cargo aircraft landed on the islands, bringing a wealth of goods. To bring back the cargo planes, islanders set up landing strips, lit fires as runway lights, and mimed communication with the oncoming planes using makeshift communication huts, wooden headsets, and the like. They went through the motions that had led to landings, without understanding the significance of those motions.

In our experience, many applications of statistics are cargo-cult statistics: practitioners go through the motions of fitting models, computing *p*-values or confidence intervals, or simulating posterior distributions. They invoke statistical terms and procedures as incantations, with scant understanding of the assumptions or relevance of the calculations, or even the meaning of the terminology. This demotes statistics from a way of thinking about evidence and avoiding self-deception to a formal “blessing” of claims. The effectiveness of cargo-cult statistics is predictably uneven. But it is effective at getting weak work published – and is even required by some journals.

The crisis in statistics is a microcosm of the crisis in science: the mechanical application

► had a prior.¹³ To proceed with a model or prior that is not chosen carefully and well grounded in disciplinary knowledge, to mix frequentist and Bayesian methods oblivious, to select the prior after looking at the data to get a result one likes, and to combine systematic and stochastic errors as if they were independent random errors are all forms of cargo-cult statistics. The calculations are as likely to produce valid inferences as cargo cults were to summon cargo planes.

Statistics education: contributory negligence

While statistical education has started a sea change for the better, in our experience, many statistics courses – especially “service” courses for non-specialists – teach cargo-cult statistics: mechanical calculations with little attention to scientific context, experimental design, assumptions and limitations of methods, or the interpretation of results.

This should not be surprising. These courses are often taught outside statistics departments by faculty whose own understanding of foundational issues is limited, having possibly taken similarly shallow courses that emphasise technique and calculation over understanding and evidence.

Service courses taught in statistics departments often have high enrolments, which help justify departmental budgets and staffing levels. Statistics departments may be under administrative, social, and financial pressure to cater to the disciplinary “consumers” of the courses. Consumers may not care whether methods are used appropriately, in part because, in their fields, the norm (including the expectations of editors and referees) is cargo-cult statistics. The bad incentives for individuals, departments, and disciplines are clear; negative consequences for science and society are expected.

Statistical software: power without wisdom

Statistical software enables and promotes cargo-cult statistics. Marketing and adoption of statistical software are driven by ease of use and the range of statistical routines the software implements. Offering complex and “modern” methods provides a competitive advantage. And some disciplines have in effect standardised on particular statistical software, often proprietary software.

Statistical software does not help you know what to compute, nor how to interpret the result. It does not offer to explain the assumptions behind methods, nor does it flag delicate or dubious assumptions. It does not warn you about multiplicity or *p*-hacking. It does not check whether you picked the hypothesis or analysis after looking at the data, nor track the number of analyses you tried before arriving at the one you sought to publish – another form of multiplicity. The more “powerful” and “user-friendly” the software is, the more it invites cargo-cult statistics.

This is hard to fix. Checks of residuals and similar tests cannot yield evidence that modelling assumptions are true – and running such checks makes the estimates and inferences conditional, which software generally does not take into account. In-built warnings could be used to remind the user of the assumptions, but these are unlikely to have much effect without serious changes to incentives. Indeed, if software offered such warnings, it might be seen as an irritant, and hence a competitive disadvantage to the vendor and the user, rather than an aid.

Scientific publishing and open science

Peer review can reinforce bad scientific and statistical practice. Indeed, journals may reject papers that use more reliable or more rigorous

methods than the discipline is accustomed to, simply because the methods are unfamiliar. Conversely, some disciplines become enthralled with methodology *du jour* without careful vetting. Even the increased volume of research suggests that quality must suffer.

There is structural moral hazard in the current scientific publishing system. Many turf battles are fought at the editorial level. Our own experience suggests that journals are reluctant to publish papers critical of work the journal published previously, or of work by scientists who are referees or editors for the journal.

Editorial control of prestigious journals confers indirect but substantial control of employment, research directions, research funding, and professional recognition. Editors and referees can keep competitors from being heard, funded, and hired. Nobel biologist Randy Shekman reports: “Young people tell me all the time, ‘If I don’t publish in CNS [a common acronym for *Cell/Nature/Science*, the most prestigious journals in biology], I won’t get a job’ ... [Those journals] have a very big influence on where science goes.”¹⁴

Editorial policies may preclude authors from providing enough information for a reviewer (or reader) to check whether the results are correct, or even to check whether the figures and tables accurately reflect the underlying data. As a result, the editorial process simply cannot perform its intended quality-control function.

Academic research is often funded, at least in part, by taxes. Yet many scientists try to become rent-seekers, keeping the data and code that results from public funding to themselves indefinitely, or until they feel they have exhausted its main value. This is morally murky. To “publish” the resulting research behind a paywall, inaccessible to the general public, is even more troubling. Scientific



publishing is big business, and its interests are not those of science or scientists.¹⁴ Open data, open software, and open publication may provide better value for society and a better ethical foundation for science.

What can statisticians do?

Statisticians can help with important, controversial issues with immediate consequences for society. We can help fight power asymmetries in the use of evidence. We can stand up for the responsible use of statistics, even when that means taking personal risks.

We should be vocally critical of cargo-cult statistics, including where study design is ignored, where *p*-values, confidence intervals and posterior distributions are misused, and where probabilities are calculated under irrelevant, misleading assumptions. We should be critical even when the abuses involve politically charged issues, such as the social cost of climate change. If an authority treats estimates based on an *ad hoc* collection of related numerical models with unknown, potentially large systematic errors as if they were a random sample from a distribution centred at the parameter, we should object – whether or not we like the conclusion.

We can insist that “service” courses foster statistical thinking, deep understanding, and appropriate scepticism, rather than promulgating cargo-cult statistics. We can help empower individuals to appraise quantitative information critically – to be informed, effective citizens of the world. We also can help educate the media, which often reduces science to “infotainment” through inaccurate, sensationalised, truncated, and uncircumspect reporting. Journalists rarely ask basic questions about experimental design or data quality, report uncertainties, or check the scientific literature for conflicting results, etc. We can address this by teaching courses for journalists and editors.

When we appraise each other’s work in academia, we can ignore impact factors, citation counts, and the like: they do not measure importance, correctness, or quality. We can pay attention to the work itself, rather than the masthead of the journal in which it appeared, the press coverage it received, or the funding that supported it. We can insist on evidence that the work is correct – on reproducibility and replicability – rather than pretend that editors and referees can reliably

vet research by proxy when the requisite evidence was not even submitted for scrutiny.

We can decline to referee manuscripts that do not include enough information to tell whether they are correct. We can commit to working reproducibly, to publishing code and data, and generally to contributing to the intellectual commons. We can point out when studies change endpoints. We can decline to serve as an editor or referee for journals that profiteer or that enable scientists to be rent-seekers by publishing “results” without the underlying publicly funded evidence: data and code.

And we can be of service. Direct involvement of statisticians on the side of citizens in societal and environmental problems can help earn the justified trust of society. For instance, statisticians helped show that the erroneous use of zip codes to identify the geographic area of interest in the Flint, Michigan water pollution scandal made the water contamination problem disappear.

Statistical election forensics has revealed electoral manipulation in countries such as Russia. Statistical “risk-limiting” audits, endorsed by the ASA, can provide assurance that election outcomes are correct. Such methods have been tested in California, Colorado, Ohio, and Denmark, and are required by law in Colorado and Rhode Island; other states have pending legislation. Statisticians and computer scientists developed methods and software; worked with election officials, legislators, and government agencies on logistics, laws, and regulations; and advocated with the public through op-eds and media appearances. Statisticians and mathematicians can help assess and combat gerrymandering, the practice of redrawing electoral districts to advantage a party unfairly.

Statisticians are pointing out biases inherent in “big data” and machine-learning approaches to social issues, such as predictive policing. They could also work with economists to monitor new forms of exploitation of intellectual labour now that new modes of working can be exploited in old ways.

We statisticians can support initiatives such as the Reproducibility Project, the Meta-research Innovation Center, the EQUATOR network, alltrials.net, retractionwatch.com, and others that aim to improve quality and ethics in science, and hold scientists accountable for sloppy, disingenuous, or fraudulent work.

And we can change how we work. We can recognise that software engineering is as

important to modern data analysis as washing test tubes is to wet chemistry: we can develop better computational hygiene. And we can ensure that publicly funded research is public.

In the 1660s, radical philosophers sought to understand and master the world by becoming scientific – creating science. In the 1970s, radical scientists sought to change the world by changing science. Perhaps that is now needed again. ■

Editor’s note

A fully referenced version of this article is available online at significancemagazine.com/593.

References

1. Shapin, S. (2008) *The Scientific Life: A Moral History of a Late Modern Vocation*. Chicago: University of Chicago Press.
2. Ravetz, J. (2009) Morals and manners in modern science. *Nature*, **457**(7230), 662–663.
3. Price, D. J. de S. (1963) *Little Science, Big Science*. New York: Columbia University Press.
4. Ravetz, J. P. (1971) *Scientific Knowledge and its Social Problems*. Oxford: Clarendon Press.
5. Mirowski, P. (2011) *Science-Mart: Privatizing American Science*. Cambridge, MA: Harvard University Press.
6. Ioannidis, J. P. A., Chen, J., Kodell, R., Haug, C. and Hoey, J. (2005) Why most published research findings are false. *PLoS Medicine*, **2**(8), e124.
7. Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J. and Ioannidis, J. P. A. (2017) A manifesto for reproducible science. *Nature Human Behaviour*, **1**(1), 0021.
8. Smaldino, P. E. and McElreath, R. (2016) The natural selection of bad science. *Royal Society Open Science*, **3**, 160384.
9. Edwards, M. A. and Roy, S. (2017) Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, **34**(1), 51–61.
10. Chesterton, G. K. (2015) *Napoleon of Notting Hill*. Open Road Media.
11. Feynman, R. P., Leighton, R. and Hutchings, E. (1985) *Surely You’re Joking, Mr. Feynman!* New York: W. W. Norton.
12. Wasserstein, R. L. and Lazar, N. A. (2016) The ASA’s statement on *p*-values: Context, process, and purpose. *American Statistician*, **70**(2), 129–133.
13. Freedman, D. (1995) Some issues in the foundation of statistics. *Foundations of Science*, **1**(1), 19–39.
14. Buranyi, S. (2017) Is the staggeringly profitable business of scientific publishing bad for science? *The Guardian*, 27 June.