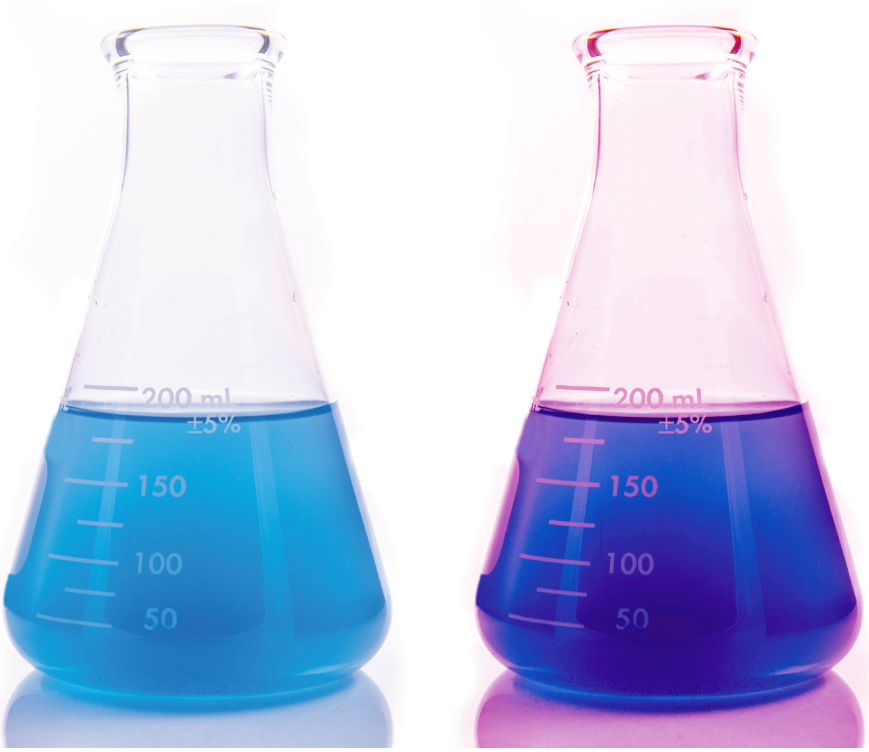
The reproducibility crisis in science

A statistical counterattack

More people have more access to data than ever before. But a comparative lack of analytical skills has resulted in scientific findings that are neither replicable nor reproducible. It is time to invest in statistics education, says **Roger Peng**



Peter_art/iStock/Thinkstock

Over the last two decades, the price of collecting a unit of data has dropped dramatically. New technologies touching every aspect of our lives – from our finances, to our health, to our social interactions – have made data collection cheap and easy. In 1967 Stanley Milgram did an experiment (bit.ly/1PWzLDy) to determine the number of degrees of separation between two people in the USA. In his experiment he sent 296 letters to people in Omaha, Nebraska, and Wichita, Kansas, and the goal was to get the letters to a specific person in Boston, Massachusetts. His experiment gave us the notion of "six degrees of separation". A 2007 study (bit.ly/1PWA2q8) updated that number to "seven degrees of separation" – except the newer study was based on 30 billion instant messaging conversations collected over 30 days.

This example illustrates a growing problem in science today: collecting data is becoming too much fun for everyone. Developing instruments, devices, and machines for generating data is fascinating, particularly in areas where little or no data previously existed. Our phones, watches, and eyeglasses all collect data. Because collecting data has become so cheap and easy, almost anyone can do it. As a result, we are all statisticians now, whether we like it or not (and judging by the looks of some of my students, many do not). All of us are regularly confronted with the problem of how to make sense of the deluge of data. Data follow us everywhere and analysing them has become essential for all kinds of decision-making. Yet, while our ability to generate data has grown dramatically, our ability to understand them has not developed at the same rate.

Making research reproducible

There are two major components to a reproducible study: that the raw data from the experiment are available; and that the statistical code and documentation to reproduce the analysis are also available. These requirements point to some of the problems at the heart of the reproducibility crisis.

First, there has been a shortage of software to reproducibly perform and communicate data analyses. Recently, there have been significant efforts to address this problem and tools such as knitr, iPython notebooks, LONI, and Galaxy have made serious progress.

Second, data from publications have not always been available for inspection and reanalysis. Substantial efforts are under way to encourage the disclosure of data in publications and to build infrastructure to support such disclosure. Recent cultural shifts in genomics and other areas have led to journals requiring data availability as a condition for publication and to centralised databases such as the US National Center for Biotechnology Information's Gene Expression Omnibus (GEO) being created for depositing data generated by publicly funded scientific experiments.

One might question whether reproducibility is a useful standard. Indeed, one can program gibberish and have it be perfectly reproducible. However, in investigations where computation plays a large part in deriving the findings, reproducibility is important because it is essentially the only thing an investigator can guarantee about a study. Replicability cannot be guaranteed – that question will ultimately be settled by other independent investigators who conduct their own studies and arrive at similar findings. Furthermore, many computational investigations are difficult to describe in traditional journal papers, and the only way to uncover what an investigator did is to look at the computer code and apply it to the data. In a time where data sets and computational analyses are growing in complexity, the need for reproducibility is similarly growing.

One result of this is an epidemic of poor data analysis, which is contributing to a crisis of *replicability* and *reproducibility* of scientific results. Replication is the cornerstone of scientific research, with consistent findings from independent investigators the primary means by which scientific evidence accumulates for or against a hypothesis. The *replicability* of a study is related to the chance that an independent experiment targeting the same scientific question will produce a result consistent with the original study. Recently, a variation of this concept, referred to as *reproducibility*, has emerged as a key minimum acceptable standard, especially for heavily computational research. Reproducibility is defined as the ability to recompute data analytic results, given an observed data set and knowledge of the data analysis pipeline. Replicability and reproducibility are two foundational characteristics of a successful scientific research enterprise.

Public failings

Yet there is increasing concern in the scientific community about the rate at which published studies are either reproducible or replicable. This concern gained significant traction with a statistical argument that suggested most published scientific results may be false positives (bit.ly/1PWAhBx). Concurrently, there have been some very public failings of reproducibility across a range of disciplines, from cancer genomics (bit.ly/1PWAC7a), to clinical medicine (bit.ly/1KNc4u6) and economics (bit.ly/1PWBngz) and the data for many publications have not been made publicly available, raising doubts about the quality of data analyses. Compounding these problems is the lack of widely available and user-friendly tools for conducting reproducible research.

Perhaps the most infamous recent example of a lack of replicability comes from Duke University, where in 2006 a group of researchers led by Anil Potti published a paper claiming that they had built an algorithm using genomic microarray data that predicted which cancer patients would respond to chemotherapy.[1] This paper drew immediate attention, with many independent investigators attempting to reproduce its results. Because the data were publicly available, two statisticians at MD Anderson Cancer Center, Keith Baggerly and Kevin Coombes, obtained the data and attempted to apply Potti *et al.*'s algorithms.[2] What they found instead was a morass of poorly conducted data analyses, with errors ranging from trivial and strange to devastating. Ultimately, Baggerly and Coombes were able to reproduce the (erroneous) analysis conducted by Potti *et al.*, but by then the damage was done. It was not until 2011 that the original study was retracted from *Nature Medicine*.

Another recent example comes from the world of economics, where an influential paper published by Carmen Reinhart and Kenneth Rogoff suggested that countries with very high debt–GDP ratios suffer from low growth.[3] In fact, they suggested that there was a "threshold" at 90% debt–GDP ratio above which there was a drop in economic growth. Thomas Herndon, a graduate student in economics, obtained the data from Reinhart and Rogoff and eventually reproduced their analysis.[4] In the process of reproducing the analysis, however, he found numerous errors. One often-quoted error was a mistake in a Microsoft Excel spreadsheet that lead to a few countries accidentally being left out of the analysis. However, a much more serious issue was an unusual form of data weighting that produced the "threshold" effect. Herndon *et al.* found that using a more standard weighting led to a smoother relationship between debt–GDP ratio and growth. Ultimately, the research on which much economic policy was based – most notably arguments in favour of economic austerity – suffered from serious but easily identifiable flaws in the data analysis.

So what went wrong with each of these studies? Clearly, many things – but reproducibility was arguably not the problem in either case. It was precisely because the analyses were reproducible that Baggerly and Coombes and Herndon *et al.* were able to identify so many errors (see box, "Making research reproducible"). Ultimately, the problem was the poor or questionable quality of the original analysis. The errors that were made showed a lack of judgement, training, or quality control. One then has to ask how these disasters could have been prevented.

Building trust

In order to improve the quality of science I believe we need to go beyond calling for mere reproducibility. The key question we

want to answer when seeing the results of any scientific study is whether we can trust the data analysis. If we think of problematic data analysis as a disease, reproducibility speeds diagnosis and treatment in the form of screening and rejection of poor data analyses by journal referees, editors, and other scientists in the community. Once a poor data analysis is discovered, it can be "treated" in various ways.

This current "medication" approach to maintaining research quality relies on peer reviewers and editors to make a diagnosis consistently. This is a tall order. Editors and peer reviewers at medical and scientific journals often lack the training and time to perform a proper evaluation of a data analysis. This problem is compounded by the fact that data sets and data analyses are becoming increasingly complex, the rate of submission to journals continues to increase (bit.ly/1PWBxVm), and the demands on statisticians to referee are increasing (bit.ly/1PWC5um). These pressures have reduced the efficacy of peer review in identifying and correcting potential false discoveries in the medical literature. And, crucially, the medication approach only addresses the problem of poor data analysis after the work has been done.

Increasing data analytic literacy comes at a potential cost. Individuals might develop the skills to perform data analysis without the knowledge to prevent mistakes

If we could prevent problematic data analyses from being conducted, we could substantially reduce the burden on the community of having to evaluate an increasingly heterogeneous and complex population of studies and research findings. To prevent poor data analysis in the scientific literature we need to increase the number of trained data analysts in the scientific community, and to identify statistical software and tools that can be demonstrated to improve reproducibility and replicability of studies and be moderately robust to user error. The US National Institutes of Health has identified data science education as a priority by issuing requests for applications for training materials, courses, and other educational initiatives focused on reproducibility. Increasing data analytic literacy has the chance of increasing the probability that any given scientific data analysis will be sensible and correct. If this is successful it will reduce the burden of detecting poor data analyses through the overtaxed peer review system and will increase the pool of trained editors and referees in the peer review process.

Education at scale

How can we dramatically scale up data science education in the short term? One example is the approach we have taken at the Johns Hopkins Bloomberg School of Public Health, where we were one of the earliest participants in the massive online open course phenomenon. Inspired by the huge demand for statistical and data science knowledge, my colleagues Jeffrey Leek, Brian Caffo, and I built the Johns Hopkins Data Science Specialization (bit.ly/1PWBZms), a sequence of nine courses covering the full spectrum of data science skills from formulating quantitative questions, to cleaning data, to statistical analysis and producing reproducible reports.

But simply increasing data analytic literacy comes at a cost. Most scientists in programmes like ours will receive basic to moderate training in data analysis, creating the potential for generating individuals with enough skill to perform data analysis but without enough knowledge to prevent data analysis mistakes.

Therefore, to improve the global robustness of scientific data analysis, we must take a two-pronged approach and couple massive-scale education efforts with the identification of data analytic strategies that are reproducible and replicable in the hands of basic or intermediate data analysts. It is critical that we make a coordinated effort to identify statistical software and standardised data analysis protocols that are shown to increase reproducibility and replicability in the hands of people with only basic training.

It is also critical that statisticians bring to bear their history of developing rigorous methods to the area of data science. One fundamental component of scaling up data science education is performing empirical studies to identify statistical methods, analysis plans, and software that lead to increased replicability and reproducibility in the hands of users with basic knowledge. We call this approach "evidence-based data analysis". Just as evidence-based medicine applies the scientific method to the practice of medicine, evidence-based data analysis applies the scientific method to the practice of data analysis. Combining massive scale education with evidence-based data analysis can allow us to quickly test data analytic practices (bit.ly/1PWCdtQ) in a population most at risk for data analytic mistakes.

In much the same way that the epidemiologist John Snow helped end a London cholera epidemic by convincing officials to remove the handle of an infected water pump, we have an opportunity to attack the crisis of scientific reproducibility at its source. Dramatic increases in data science education, coupled with robust evidence-based data analysis practices, have the potential to prevent problems with reproducibility and replication before they can cause permanent damage to the credibility of science.

References

1. Potti, A., Dressman, H. K., Bild, A., *et al.* (2006) Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, **12**, 1294–1300.
2. Baggerly, K. A. and Coombes, K. R. (2009) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, **3**, 1309–1344.
3. Reinhart, C. M. and Rogoff, K. S. (2010) Growth in a time of debt. *American Economic Review*, **100**, 573–578.
4. Herndon, T., Ash, M. and Pollin, R. (2013) *Does High Profile Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff*. Political Economy Research Institute Working Paper Series, no. 322. PERI, University of Massachusetts, Amherst.

Roger D. Peng is an associate professor of biostatistics at the Johns Hopkins Bloomberg School of Public Health