The Economist

Scientific honesty

Computer algorithms can test the dodginess of published results

A potential boon for journal editors

Print edition | Science and technology Jun 16th 2018

IN AN ideal world the data on which a scientific study is based should be, if not publicly available, then at least available to other researchers with a legitimate interest in asking. Sadly, this is not always the case. Though attitudes are changing, many scientists are still quite proprietorial about their data. They collected them, they reason, and thus they own them, and with them the right to analyse them without sharing them with rivals.

This attitude, though selfish, is understandable. But sometimes it can cover a darker secret. The statistics presented in a paper may have been manipulated to achieve a desired result. The author may, in other words, have cheated. If he releases the data, that cheating will be obvious. Better to keep them hidden.

That, though, will be harder in the future—at least for sets of data that consist of integer numbers in a known range, as do, for example, the answers to many questionnaires in psychology experiments. As they describe in a paper in *PsyArXiv Preprints*, Sean Wilner and his colleagues at the University of Illinois at Urbana-Champaign have come up with a way of reconstructing, given the mean, standard deviation and number of data points in a result (all three of which are usually stated as part of such a result), all the possible data sets which could have given rise to that result.

They call the resulting algorithm CORVIDS (Complete Recovery of Values in Diophantine Systems). If CORVIDS cannot come up with a valid set of data for a result, that result is self-evidently fishy. If it can reconstruct a valid data set or sets, then the team running it can look at them and assess whether or not they look plausible.

The trick behind CORVIDS is to find all possible combinations of numbers that solve the linear equations from which the statistics being examined are calculated. The principle of how to do this was worked out in the third century AD by Diophantus of Alexandria (hence the "D" in "CORVIDS"). Diophantus did not, however, have access to computers and so could not take the idea very far. Mr Wilner does, and has.

To simplify the task of spotting anomalies, CORVIDS turns the possible data sets into histograms and arranges them into a three-dimensional chart. This makes any unusual patterns apparent. For example, every reconstructed data set may be missing values at one end of the scale. That might make sense occasionally. Generally, though, such a gap would be a red flag. It would suggest either that the statistics were reported incorrectly or that there were problems with the underlying data. Such problems might be caused by anything from biased methods of data collection to outright fabrication.

CORVIDS is likely to be of immediate value to editors and reviewers at academic journals, who will be able to spot problems with submitted papers early, and so discuss them with the authors. That will often be easier than asking for every paper to be accompanied by its data and then reworking the statistics from those. If an unresolvable problem does show up then the technique can be applied to previous work by the author in question, to see if anything systematic is going on.

A drawback of CORVIDS is that in some circumstances it may take hours to run. But another recently published algorithm is not beset by this problem. SPRITE (Sample Parameter Reconstruction via Iterative Techniques) was described last month in *PeerJ Preprints* by James Heathers of Northeastern University, in Boston. SPRITE is a "heuristic search algorithm"—meaning that it may not nab all possible solutions. But its speed makes it a useful first step. If the data sets it finds do not show any strange patterns, CORVIDS is unlikely to show oddities either.

Sloppy reporting of statistics in research papers is widespread. How common made-up studies are is anyone's guess. With CORVIDS, SPRITE and their kind around, though, there will soon be no hiding place for such failings—and the trustworthiness of scientific papers will take a step up.

This article appeared in the Science and technology section of the print edition under the headline "Something to crow about"