ISSN 1744-1986

T e c h n i c a l R e p o r t N º 2012/ 02

The Problem of Reproducibility

Darrel Ince

02 February, 2012

Department of Computing
Faculty of Mathematics, Computing and Technology
The Open University

Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

http://computing.open.ac.uk

The Problem of Reproducibility

Introduction

In this article, I look at the issue of reproducibility in science, concentrating on omics- and medicine-based research work. In a sense I am (to quote a British saying) 'teaching my grandmother to suck eggs' since many of you, as statisticians, will have this as your central tenet. However, I hope I might describe some ideas and materials that are new to you—at least partially. Ideas which you might like to share with non-statisticians.

A central tenet of computing is Moore's law. As expounded by Gordon Moore, co-founder of Intel, it states that the power of a computer doubles every two years. When anyone wants to talk about the advances that computers have made possible they inevitably quote this law. What is often forgotten, though, is that other factors have also been important: the increasing miniaturisation of components (RFID computers have even been attached to ants) and sturdiness (computers are attached to the hulls of ships in order to measure ocean temperatures). This has given rise to a major data explosion.

This has provided huge opportunities for researchers but, at the same time, made their job much more difficult in terms of validation. The aim of this article is to discuss some of these difficulties and to point the way forward. It will look at sources of error in research: data errors, hardware errors, statistical errors, software errors and conclude with some suggestions about making research work using the computer much more reproducible. First, the problems with data.

Problems with Data

A good starting point with data problems is a case study of recent research that was carried out at Duke University that proved flawed.

In 2006 a ground breaking article was published by researchers at Duke. It claimed that by using genomic data it was possible to predict the response of cancer sufferers to various chemotherapeutic regimes. This was a potentially a very important contribution to an area of health known as personalized medicine. Researchers in this area have, as their main aim the development of techniques and technologies that can be used to direct therapies and interventions based on the genetic makeup of a patient. The apparent success of the Duke

research would have had major rewards for medicine, for the researchers and for Duke University.

The work came to the attention of oncologists worldwide. Some of these were based at the UT M.D. Anderson Cancer Center in Houston. They asked two statisticians, Keith Baggerly and Kevin Coombes, to investigate prior to deploying the results in clinical trials. Almost from the moment they started both found major difficulties in reproducing the results. Even after a period of consulting the Duke University researchers they still had major failures in reproducing results.

Baggerly and Coombes' investigation took almost three years before the research was exposed as flawed and a number of articles in gold-standard journals were retracted. It took the two statisticians over 200 days to carry out their analysis because of the way that the research was packaged up and versioned; in effect it was difficult to reproduce and discover that there were a number of problems with the data. A good account of the full story of the Duke research can be found in a 2011 edition of the *Economist* (http://www.economist.com/node/21528593).

It was an incident that caused huge concern, particularly since clinical trials had started based on the results of the research. One of the positive outcomes was that it made the issue of reproducibility in medical research much more public and has prompted an ongoing Institute of Medicine inquiry into the relationship between omics based research and clinical trials.

A major problem with the Duke work was that the curation and packaging of data was such that it took a long time for Baggerly and Coombes to decipher what was going on. They had to carry out a process they refer to as forensic bioinformatics where they worked backwards from final and intermediate results to establish what the researchers had done. There were, in effect, major problems with reproducibility.

The intrusion of the computer within emerging areas of research is particularly worrying in terms of areas such as genomics. John Ioannidis is a man who I hugely admire for his stance on the responsibilities of science and scientists. In late 2011 he published an excellent article in *Science* with Muin Khoury that looked at the whole issue of validation in omics based research (genomics, transcriptomics, proteomics, metabolomics and others). This is research that promises therapies and interventions that are based on the genetic makeup of a patient. However, it is an area that has billions of items of potentially error-prone data generated under different lab conditions by different technologies.

This paper is one of the best descriptions of problems that beset researchers that use the computer as a central tool in their work . Although it is oriented towards omics research the message is relevant to other areas where the computer is a vital supporting technology. It also contains a number of suggestions to overcome the sort of problems that arose from the Duke work. I will examine some of these suggestions later in this article.

Ioannidis has a connection with the Duke problems in that he published a prescient paper in *Nature Genetics* about the data produced by the microarray technology used by the Duke researchers; it was published during the period when Baggerly and Coombes were carrying out their forensic investigation. He and his colleagues looked at a number of microarray studies and discovered major problems with reproducibility.

Before looking at some ways that scientists and statisticians can package up their work it is worth looking at another reason reproducibility is important: possible errors arising from statistical defects.

Statistical Problems

The statistical literature contains a corpus of papers that examine the problems associated with medical research. In 2007 an important article was published by Alexander Strasak and colleagues from the Medical University in Innsbruck. In it they looked at errors, flaws and deficiencies in medical statistics. Their article brings together and reviews a collection of works that have appeared in major medical and statistical journals that critically examine the quality of medical statistics. This is a very important paper for two reasons. First it is a comprehensive review of statistical problems in medical research and, second and equally importantly, it contains an implicit checklist that any researcher who publishes can use to evaluate their statistical work.

Strasak and his colleagues looked at five areas of statistics in medicine: statistical design, data analysis, method documentation, presentation of statistical data and interpretation. Each area was covered in terms of advice to the medical statistician. Some of their criticisms, taken from the 47 statistical errors and shortcomings the authors documented, were: use of an inappropriate test for a hypothesis, no Yates continuity correction reported for small numbers, a failure to use randomisation, using an inappropriate control group, a failure to state the number of tails, the use of the mean to describe non-normal data and drawing conclusions not supported by the data.

These are just a subset of the problems that Strasak and his colleagues documented, they constitute an alarming review of problems with statistics in medicine. The whole paper is worth a careful read by anyone involved in medical statistics.

If my article stopped at this point then it should raise lots of alarm bells. However, there are more problems that researchers face over and above those of data curation, data validation and poor statistical practice. There are problems with computation. These are both hardware and software related.

Problems with Computation

One area that is capable of generating error is that of floating point computation. A computer stores integers exactly; however, it stores floating point numbers in an approximate way. Here is a statement that many of you might be surprised at

> More subtly, on some platforms, the exact same expression, with the same values in the same variables, and the same compiler, can be evaluated to different results, depending on seemingly irrelevant statements (printing debugging information or other constructs that do not openly change the values of variables).

It is taken from a major review of floating point problems and solutions by David Monniaux. What this says is staggering: that if you have a programming statement that involves floating point variables and constants then the result of that statement may be different depending on program code that came before *that did not change* any of the values.

This is quite a shocking statement; it is known as an order of evaluation problem and many programming languages are subject to its ways. It arises from the way that a programming language compiler is provided with more flexibility in its optimisation strategy. Another problem is that floating point results can differ depending on the computer hardware and software that is used.

There are also other problems: a major one is that the way that floating point numbers are stored on a computer gives rise to error, particularly if those numbers are subject to large amounts of repetitive processing.

Many of the problems with floating point computation have, at best, only been solved partially. What is particularly worrying is that the increasing power of the computer is enabling scientists to carry out more intensive processing with, for example, smaller grids, leaving researchers into the hardware and software issues of numerical error (a relatively small community, publishing in journals not normally accessed by scientific and medical researchers) working hard to catch up.

Problems with Software Development

There are also further problems with computer-based science. Many scientists carry out software development as a peripheral activity to their main work of advancing research hypotheses, gathering data and validating the hypotheses. Consequently their software skills are less than those who carry out systems development for a living. There have been many studies of software developed by commercial companies which show errors in their software—many of them latent —only being discovered after many thousands of executions. The general figure given is that in industrial software there are usually between 1 and 10 errors per thousand lines of code. This is software produced by industrial staff working in teams for companies who have extensive quality assurance procedures, not by scientists working by themselves within a research environment.

Solutions

So science is subject to statistical errors, computational errors, programming errors and defects in data curation. Surprisingly, perhaps, I don't find this particularly depressing: research is a low yield process and errors have always been committed, even in the days when the computer was something of a curiosity in scientific research. What *is* troubling is that detecting these errors is getting much more difficult. In the past, poor research could be spotted by other researchers with not a huge investment of effort. Now that the computer has interspersed itself between the data and the results in a non-transparent way we seem to have real problems in validation; we are in effect seeing a new era where scientific intuition is much less useful. So, what should the research community do to improve current practices and enable reproducibility?

It is clear that we should all champion reproducibility; a glance at the major philosophers of science shows that none of them deviate from the view of the eminent philosopher Karl Popper, that the result of a scientific experiment should stand until it is falsified and that the longer it is unchallenged the more confident we should be of it. That once a theory has been established it stays at the party until it is either falsified or it is modified, for example by making it more specific or by excluding or modifying some of its generalities.

However even if you regard Popper as some intellectual party-pooper I hope that my discussion of statistical error and problems with data curation, programming and floating point computation convinces you that much more care is needed—even in this era of evidence-based medicine.

So how do you go about making your work reproducible? There are some very bright spots. For example there is a system called *Sweave*. This packages up data files, programs written in the programming language R and article text expressed in the document markup language Latex. It enables a researcher who receives the package to reproduce the results and even carry out further experiments such as modifying some of the data or making changes to the R programs.

If you are not an R programmer then you would be unable to use *Sweave*. However, this does not mean that you should give up on reproducibility. There are a number of valuable aspects of packaging that can be implemented in a barefoot way (I am assuming that the work involves some programming, but you can modify my suggestions if you are using some statistical package):

- Comment your program code or script with cross-references to the parts of the research article that it is relevant to. Also provide comments in each code module as to what that module does: what data it reads, what it does with that data and then what output it produces.
- Provide a document that describes the data that you are manipulating. The data in this document—known as metadata—would describe what each element of your data was, what its format was, what is its accuracy and assign a name to it which could be cross referenced in your stats code.
- Provide all the scripts that you used to generate figures together with a description of the data that the scripts processed. For example provide all your Gnuplot files. Cross reference these scripts to the output of your programs and to figures in your research articles.
- Provide a listing of all the files names and what data they contain (for this you will need to reference the metadata).
- If you have developed versions of your code and data, for example when you discovered an error or developed a new version that might carry out extra functionality compared with the version reported in a research article, then always package up previous versions with a description of the difference between it and the previous version. *Don't delete previous versions*. Researchers are often prone to improve their software after publication and hence it becomes out of synchronisation with the contents of an article.
- Provide all the tests that you have programmed that generate results in a publication together with a script that executes the tests.

These are processes that have been used in commercial software development over the last thirty years and have lead to major improvements in quality. They are often used by developers of open-source software. The computer has impinged on scientific research so much that they should be regarded as mandatory.

My belief is that science is heading for a crisis point with respect to reproducibility. What should be done? There are a number of solutions, some of which are expounded by Ioannidis and Khoury for omics research in their *Science* article, but which have general applicability:

- Data, protocol and software deposition should be made mandatory for publication.
- Data, protocol and software deposition should be made mandatory for receiving grant funding. If a previous grant shows little adherence to this then further funding should not be considered.

- A grant funding application for a project whose results depend on the computer should, as a matter of course, be accompanied by a reproducibility plan.
- Funding or other agencies should outsource targeted reproducibility checks to bioinformatics experts, either for studies that have potential for high clinical impact and/or for those small proportion of studies that are to be translated into clinical trials.
- More units should be established in higher education that have reproducibility at their core. Most universities have computer units that provide advice on software issues and many statistics department offer advice to other departments. However, more integration is required. This is vital. To expect scientists to carry out science and to act as top quality statisticians and software developers is, in the age of the Internet, grossly unrealistic.
- The funding of the development of more tools that enable the packaging of research materials up in such a way that repeating an experiment is made a much easier process.

It not only needs agreement, but also hard work and a positive commitment.

Some Big Problems

I believe that at the heart of reproducibility there are two problems that none of the suggestions above solve. The first is that over the last thirty years universities have been transformed from gentle, liberal institutions to businesses. It is inevitable and I believe that it has brought a number of benefits, for example a greater care for students. However, it has promoted a culture where publication of new results is almost everything and where labour-intensive activities such as carrying out validation studies are rated less highly. It has also promoted a culture where there is pressure on researchers to produce research quickly and neglect the packaging aspects. Change these and we should see major improvements. It is, however, a tough task indeed.

If anyone is interested in looking at the problems of data explosion and some of the solutions, then the book *The Fourth Paradigm: Data Intensive Scientific Discovery* is an excellent description of the problems and solutions associated with reproducibility in the new world of e-science. It was sponsored by Microsoft and can be downloaded for free from one of their web sites:

http://research.microsoft.com/en-us/collaboration/fourthparadigm/

My thanks to Keith Baggerly for reading and commenting on an early version of this article.

References

Ioannidis, J. P.A. and Khoury, M. J. Improving validation practices in “omics” research. *Science*. **334**, 1230-1232. 2011.

Strasak, A. M., Zaman, Q., Pfeiffer, K.P. , Gőbel, G. and Ulmer, H. Statistical errors in medical research—a review of common pitfalls. *Swiss Med. Wkly*. **137**, 44-49. 2007.