Data, science, and new computing technology

John Maindonald
Centre for Bioinformation Science, Australian National University, Canberra ACT 0200 Australia.

New technology offers new possibilities for supplementing the limited insight available from the printed version of a paper, into the data on which it is based. It should be standard practice to use web-based resources to provide a more complete account of analyses and of the data. Data should be placed in a public archive. Review and re-evaluation, including possible re-analysis of the data, should be seen as an ongoing process that continues after formal publication. The discussion has a particular focus on applied biological research.

Historical and modern influences on the formal processes of science

The formal processes that have developed within the scientific community for the maintenance of scientific traditions, and for the dissemination of scientific results, have been shaped as much by historical influences as by rational consideration. Printed journal pages became the main vehicle for the dissemination of results, with resultant severe limitations on content. New computer technology has removed the former limitations. This requires a rethinking of the total content of published papers, having in mind that this content will be divided between the print medium or its web-based equivalent and supplementary material that is placed on the web.

Changes that have implications for the statistical analysis results that appear in published papers include:

- Large datasets that have been created by automation of data collection, and by the merging of existing databases, bring new challenges. The challenge may be to obtain forms of data summary that are suitable for analysis, and/or to handle the sheer bulk of the data. Or, as in the analysis of genomic expression array or other data where the number of outcome measures is large, the data may require substantial adaptation of existing analysis methods.
- There are new types of data, derived for example from documents, images and web pages.
- New data analysis methodologies often allow analyses that make better use of the data, more directly attuned to the questions of scientific interest, than was readily possible 15 years ago.
- Advances in statistical methodology have widened the gap between those application area specialists whose statistical knowledge has not advanced much in the past decade, and those professionals who are fully *au fait* with modern methods.
- New statistical technologies that combine data from multiple studies in a single analysis may allow the detection of patterns that were not apparent from the individual studies. They may resolve apparent discrepancies between results from the separate analyses.
- Papers can be supplemented by web-based information for which space was not found, or which it was not appropriate to include, in the printed paper.

Data mining has been used as a general name for activities that come, broadly, under one or more of the first three categories noted. To an extent, it has brought a computer science perspective to the tackling of statistical problems. At the same time, there has often been inadequate attention to statistical perspectives. See Maindonald (1999, 2003) for further commentary.

Some further comments on advances in statistical methodology are in order. Many of the analyses in Maindonald & Braun (2003) would, in an earlier decade, have required high levels of statistical and computing skill, and would have been inordinately time-consuming. The R scientific computing and graphics system, distributed as freeware, is currently a favoured environment for the computer implementation of much of this new methodology (Ihaka & Gentleman 1996). To obtain the R system, go to http://cran-r.project.org

In this paper I have limited my brief to issues that relate to the use of data, and to statistical analysis but clearly, there are wider implications for publication and for other aspects of the scientific process.

Data and data analysis

Data analysis has long been an area of difficulty for many journal editors. Until the early twentieth century, there were a relatively small number of types of problem where professionals could do much to improve on relatively informal and *ad hoc* approaches to analysis. Witness the summaries of differences between crosses and self-fertilised plants, provided for Darwin by his biometrically-minded cousin Francis Galton, that appear in Darwin (1876). By modern standards, these are unsatisfactory, as pointed out in Fisher (1935).

Following a variety of teaching and lecturing jobs, **John Maindonald** was for 14 years Officer-in-Charge of the former DSIR Applied Mathematics Division substation at Mount Albert, primarily working with scientists in entomology and horticulture. In 1992, with the demise of DSIR, he moved to HortResearch. He moved to Australia in 1996, taking up a position at the Australian National University (ANU) in 1998. He has relished the stimulus and intellectual challenge of his work at ANU, bringing involvement with biologists, ecologists, public health researchers, molecular biologists, demographers, computer scientists, numerical analysts, machine learners, an economic historian and forensic linguists, as well as a very lively group of statisticians. In 2001, in a late career change of work focus, he moved to the ANU Centre for Bioinformation Science. He is the author of a book on Statistical Computation, and the senior author of a recent (2003) book, now into its third printing, that is an example-based exposition of practical approaches to data analysis.

Much of the early development of statistical methodology led to methods that, within a limited area of application, could be mastered and used by scientists whose formal statistical training was limited to one or two courses that had been taken as part of a biology degree. Such methods still have their place, but because of changes in scientific requirements and advances in statistical methodology, that place is more limited than heretofore. I will comment later on the handling, in published papers, of analyses that require methods that that are unlikely to be covered in elementary statistics courses.

The attempt to train scientists to do their own analyses was, depending on the adequacy of the training and on the nature of the problems tackled, never without problems. See for example Maindonald & Cox (1988) and Maindonald (1992). The first of these papers reviewed the statistical content of all papers that had appeared, over a two-year period, in two DSIR (as they were then) journals that related mostly to agriculture and crop science, areas where most scientists had ready access to professional statistical support. Problems with the analysis of data and presentation of results were predominantly at the less serious end of the spectrum. In areas where statistical support is less adequate, problems are likely to have been much more serious.

In the past decade or more, statistics has been pushed out of many university biology courses, in favour of courses in molecular biology. Ironically, recent developments in molecular biology have brought new demands for the use of statistical methodology. As evidence of the serious problems that are apparent in many of the statistical analyses that are now appearing in the biology literature, see Ioannidis (2005) and Michiels *et al*. (2005).

Where there are serious problems with papers, any or all of the following may apply:

- There may be assumptions that are clearly wrong.
- The analysis may not address the questions that are of scientific interest.
- Explanations of methodology may be absent or seriously deficient.
- There may be an obsessive dedication to a particular methodology that is applied whether or not it is relevant to the problem in hand.
- The paper, and the defences that authors present for the methods used, may betray serious misunderstandings of statistical theory.
- A final, less serious, problem is that analyses may use a clumsy and unwieldy methodology where better and more insightful methods are now available.

Maindonald (1992) includes a summary of issues encountered in the statistical review of papers for the New Zealand Journal of Agricultural Science. Again, I am not aware of other published formal summaries of issues that have arisen in statistical refereeing, whether for New Zealand or other journals.

Matching the skills to the task

The comments that follow reflect my statistical refereeing experience in cases when the analysis has required methods that are not commonly taught in elementary courses. In such cases, authors who have not handled the analysis appropriately first time have rarely, without substantial professional statistical help, come up with a satisfactory analysis in later iterations of the reviewing process. The research may have extended over months or years, tackling important problems. Further effort has been expended on extensive and time-consuming statistical analyses, leading however to analyses that are seriously flawed. The waste and misdirection of effort is therefore serious.

In the medium to long term, the answer is to ensure that, for research that demands it, an appropriately experienced statistician is included in the research team. Unfortunately qualified statisticians of any kind, and especially experienced statisticians, are currently in short supply. In New Zealand, openings for the nurturing of recent statistical graduates were much reduced following the disbanding of DSIR Applied Mathematics Division.

Journal editors, forced to choose between rejecting the paper and the data on which it is based, and accepting a paper where there are serious problems with the statistical analysis and perhaps also with the data summaries and graphs, do however require a short-term solution. Separation of the total publication task into two parts would often make sense. A short paper can be prepared that describes the project and the resulting data, with an undertaking to make the data available on a web site within some reasonable period of time, whether or not the authors have by then provided a satisfactory statistical analysis. Below, I will go further, arguing that the publication of data ought to be standard practice.

The skills required to execute a science project are different from those required to analyse the data, requiring some separation of responsibilities. At the same time, both the design of data collection and the analysis require a marrying of statistical expertise with application area expertise. The best way to resolve these conflicting demands is to make professionally trained statisticians part of the project team, working closely with scientists at all stages through from project design to data analysis and publication.

A cautionary tale

In July 2005, the British newspapers reported extensively on the striking from the medical register of the senior and respected paediatrician Sir Roy Meadow. A particular concern was the inappropriate multiplication of two probabilities, themselves gross underestimates of the relevant probability of a cot death, to obtain the probability that two children in the one family would die from natural cot death causes. This misuse of statistical evidence was almost certainly a factor in the conviction and imprisonment of at least one of several women who were later acquitted. It is unusual for experts to experience the ignominy that Sir Roy Meadow has now attracted. Nevertheless, whether giving evidence in court or writing scientific papers, scientists should be careful not to stray outside of their area of competence. For a discussion of this case, see http://en.wikipedia.org/wiki/Roy_Meadow

Getting value from data

By no means do scientific papers necessarily extract all useful information from the data that have been analysed and summarised. Some possibilities are:

- There may be benefit from exploring alternative analyses of the same data. This is especially relevant if the analysis given by the authors has been less than optimal, or there is a suspicion that it has misrepresented the data.
- There may be ambiguities in the authors' description of their analysis, which inspection of the data can resolve.
- There may be information in the data that bears on issues that were not of immediate interest to the authors, or which were not noticed by the authors.
- The data may hold information, not available from the published summaries, needed for the design of a new study.
- Researchers who are planning a new related study can use existing published data in a practice run of the analysis.
- Often, data from different studies can and should be available for bringing together into a single analysis.

Together, these create an overwhelming case for the routine publication of data that form the basis of published papers. Until recently, mandatory archiving of data following the publication of a paper has been unusual. Currently, the practice is limited to a few journals. Such an initiative, even if initially resisted by some authors, has obvious benefits for the scientific process. Science has a cumulative character, where new research and data should both criticise and build on what has gone before. Computer technology has removed the former excuses for not archiving data. Authors who cannot or do not wish to provide their data would be required to plead their special circumstances.

Such a practice will facilitate continuing review after a paper has been published. It extends opportunities for the data to challenge the perceptions that guided their collection, and that may have overly influenced the published analysis.

General issues that relate to the preserving and sharing of data are canvassed at length in Beedham *et al.* (2002). This report merits careful reading.

All the relevant data?

There are other issues of this same general type. Registration of medical clinical trials at the time of their initiation, which ought long ago to have been mandatory, is finally beginning to get wide attention from funding authorities (Staessen & Bianchi 2003). The common failure to publish results that do not show an effect has the consequence that published studies present a biased picture of the total research. There are other areas of science, also, where such prior registration would be helpful to the scientific process.

Just what analysis was done?

Code that is used for analyses, for whatever software system has been used, should be available on a web page. This removes any doubt as to what analysis has been performed, and makes it possible for readers to try variations on the analysis. Use of a Graphic User Interface (GUI)-based system will be possible only if the point and click commands generate a script that can be used to reproduce the analysis. Insistence on such an audit trail is surely a reasonable requirement for published results.

It is possible to go further, allowing readers who have the necessary resources to reproduce all output results, graphs and tables that appear in a scientific paper. Using LaTeX markup conventions for text, and Sweave markup conventions for code, the text is wrapped up with computer code that may produce any or all of printed output, tables, graphs and results that are to appear in the text. This Sweave version of the paper is then processed though a function in the R system, generating a LaTeX version of the total manuscript. This LaTeX version is then, finally, processed through the LaTeX typesetting system to give a pdf or postscript file. Provision of an Sweave version of the paper makes it straightforward to investigate the effect of changes to the analysis and/or to the data. Gentleman and Lang (2004) argue strongly for the use of this approach, as a matter of standard practice, in papers that present analyses of genomic data.

Conclusion

This paper has emphasised the key role of statistical analysis in many of the papers that appear in applied biological journals. Wherever possible, data should be available on a web page, along with the code used for analysis. This is a necessary step in making research processes as transparent and open as possible.

References

Beedham, H.; Anderson, S.; Denman, J.; Macfarlane, A.; Ashley, K.; Healey, S.; Walker, N.; Blake, R.; Hunter, P.; Westlake, A.; Corti, L.; Law, H. 2002. *Preserving and Sharing Statistical Material*. University of Essex and the Royal Statistical Society, 52 p.

Darwin, C. 1892. *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. Appleton and Company, first published by John Murray,1876. http://www.gutenberg.org/etext/4346

Fisher, R.A. 1935 (7th edn, 1960). *The Design of Experiments*. Oliver and Boyd, London.

Gentleman, R.; Lang, D.T. 2004. Statistical Analyses and Reproducible Research. *Bioconductor Project Working Papers 2*. http://www.bepress.com/bioconductor/paper2

Ihaka, R.; Gentleman, R. 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics 5*: 299–314.

Ioannidis, J.P.A. 2005. Microarrays and molecular research: noise discovery? *Lancet 365*: 454.

Maindonald, J.H. 1992. Statistical design, analysis and presentation issues. *New Zealand Journal of Agricultural Research 35*: 121–141.

Maindonald, J.H. 1999. New approaches to using scientific data – statistics, data mining and related technologies in research and research training. *ANU Graduate School Occasional Paper 98/2*. v + 37 p. http://www.anu.edu.au/graduate/pubs/occasional_papers/GS98/2

Maindonald, J.H. 2003. The role of models in predictive validation. Invited Paper, ISI Meeting, Berlin, 2003. [RFCD 230204]

Maindonald, J.H.; Braun, J.B. 2003. *Data Analysis & Graphics Using R. An Example-Based Approach*. Cambridge University Press, Cambridge. http://wwwmaths.anu.edu.au/~johnm/r-book.html

Maindonald, J.H.; Cox, N.R. 1984. Use of statistical evidence in some recent issues of DSIR agricultural journals. *New Zealand Journal of Agricultural Research 27*: 597–610.

Michiels, S.; Koscielny, S.; Hill, C. 2005. Prediction of outcome with microarrays: a multiple random validation strategy. *Lancet 365*: 488–492.

Staessen, J.A.; Bianchi, G. 2003. Registration of trials and protocols. *Lancet 365*:1009.