COMPUTATION AND BIOMETRY

J.H. Maindonald
Centre for Bioinformation Science, Australian National University, Australia

Keywords: Computing, Statistical analysis, Biometry, Expert system, Computer language, Database, Statistical package, Document preparation, Markup system, Study design.

Contents

Summary

Advances in computer technology have dramatically changed the analyses that are readily possible, the style of data analysis, and the computing environment in which analyses are performed. The improvement of the interface between statistical analysis systems and other computing systems is a huge continuing challenge. The development of "statistical adviser" systems is an even greater challenge. These would advise on and encourage good statistical computing practice, at each step of an analysis. Their advice should extend to the design of data collection, and to the presentation of results.

1. Introduction

1.1. Scope of This Chapter

This chapter will note ways in which advances in computer technology have dramatically changed both the environment in which analyses are performed and the practice of data analysis. The issues that it discusses have wide relevance across all areas of statistical application. The discussion will be slanted towards biometry, but is not limited to biometrical applications.

1.2. Statistics in the Pre-Computer Era

Statistical methodology has diverse origins, reflecting the diverse applications that have motivated its development. Major pre-twentieth century influences included the collection of information for official purposes, the development of life tables for actuarial calculations, the reconciliation of astronomical data, gambling problems, and public health. Applications that came into prominence in the twentieth century include agriculture and biology, psychology, education, industry, geophysics, medicine, climatology, business and commerce, finance and, most recently, molecular biology and genomics. New application areas have repeatedly brought new demands, calling for the development of new methodology or the adaptation of existing methodology.

The demands of agriculture and biology were the stimulus for the framework that R.A. Fisher laid for modern statistical theory in the early part of the twentieth century. The methodologies have been taken up, and developed and extended, in every area of science.

Any new methodology requires careful and continuing critical evaluation. Statistical techniques have too often ossified into mechanical rules of thumb that may be used without due critical evaluation and discretion. Unsatisfactory practices include an exaggerated emphasis on tests of hypotheses, a neglect of pattern, the policy of some journal editors of publishing only those studies which show a statistically significant effect, and an undue focus on the individual study. In the medium to long term, it is necessary for the integrity of science that the demands of scientific rationality should win out over such influences, which arise largely from accidents of historical development.

1.3. Prospects for the Short-Term Future

Initially, computers were primarily used to automate calculations that had formerly used mechanical calculators, doing much the same analyses as before. Now, in 2004, the impact has been far-reaching. While advances are most obvious in the work and working environment of skilled professionals, they will inevitably, in due course, affect all who are engaged in serious statistical analysis. Each successive methodological advance sets the stage for further development that builds on what has been achieved earlier.

2. Computer Language and Systems – Past, Present and Future

2.1. The Beginnings of Scientific Computing

Early electronic digital computers were programmed using hexadecimal codes. These were soon replaced by assembly language codes that were mnemonic equivalents of the hexadecimal codes. The development of Fortran over the period 1954-1958 was a watershed for scientific computing. In the two decades that followed, systems were developed that automated many data manipulation and sorting tasks, doing them at what seemed, relative to manual methods, lightning speeds. Initially, the major emphasis was on the development of routines that were efficient implementations of existing manual calculations. Input, usually from punched cards or paper tape, remained time-consuming, tedious and error-prone.

Experimentation with the use of electronic digital computers for statistical analysis started in the late 1950s. By 1970,"statistical packages" were gaining acceptance. These packaged a number of routines together, usually with a primitive form of command language that allowed those who were content to work within the limits of the package to avoid recourse to Fortran.

2.2. High Level versus Low Level Languages

At least until the 1980s, it was common to call Fortran, C and similar languages *high level*, in contrast with machine and assembly language. Use of assembly language is now unusual. Languages in the style of Fortran and C are more appropriately called low-level languages, and the term high level is now more appropriately reserved for languages in the style of S, GAUSS, MATLAB and Python. A broad distinction is that modern high level languages try to minimize the time that human users will take to code problems, while low level languages put a greater emphasis on computational efficiency and the minimizing of storage requirements. High level languages often focus on specific types of application. The S language, because it was designed with data analysis and graphics in mind, has been of particular interest to statisticians. There are two major implementations of the S language – the commercial S-PLUS system, and the free R system which is distributed under the General Public Licence. The languages noted are a small selection of the huge variety of languages that are available. Other widely used languages are those provided by the SAS, SPSS, Stata and Genstat systems.

The use of higher level languages has changed the style and efficiency of coding and made computation more efficient. Changes that are in the long run more significant have arisen from the demand for linkages into other software, and with linkages into other systems.

2.3. Unfulfilled Promises

There are lessons from unfulfilled promises from the first several decades of commercial digital computing. Up until perhaps 1980, there was widespread optimism that it would soon be possible, in many areas of professional activity, to build expert systems that would largely fill the role of human experts. While achievements to date and expected in the medium term future do not fulfill these promises, there have been achievements that have in many respects been more interesting, with the pervasive effect of the internet the most obvious example. While the medium term future may bring similarly unanticipated new developments, changes that bypass large investments in currently available software systems are unlikely.

Statistical expert systems were much discussed in the 1970s, and there were several substantial efforts at building such systems. These now, for the time being, seem off the agenda. Too much preliminary work remains to be done on building and integrating the component systems that would be needed as a basis for any really effective expert system. We still await systems that have the ability, in dialogue with human experts, to learn in the way that seems needed if genuinely expert systems are to become a reality.

2.4. Constraints on Future Development

There is now a huge investment in software systems that have been built up over the past several decades. Most new initiatives will take advantage of this existing resource. This constrains future development, forcing steady evolutionary progress and making radical innovation expensive and difficult and allowing limited cautious comment on likely directions of further development of biometrical computing! It should be noted that constraints that result from present large software system investments apply more to the software that handles the actual computation than to the human interface. The human interface can change quite dramatically, while using the same software to handle the major parts of the computation.

3. Changing Views of Statistical Computing

3.1. Numerical Statistical Computation

Numerical statistical computation remains central to statistical computing, notwithstanding major changes in the wider statistical computing context. Unless the numerical computations are correct, are relevant to the problem in hand, and are handled with acceptable efficiency, improvements to the environment in which these calculations are performed are pointless. Occasional serious failures in software for numerical statistical computation, or in the manner of its use, emphasize the importance of getting this part of the task right. A recent example, explored in a 2002 paper by Dominici and others, comes from studies of the effects of air pollution on mortality.

Because of numerical convergence problems, some parameter estimates were too large by a factor of two.

3.2. Changes in Methodology

Many computations are now commonplace that in the era of hand calculators were unimaginable. This has led to a rethinking of many of the approaches that were used formerly. There has been ongoing review and improvement on the makeshift methods that were often necessary when hand calculators were used, or when statistical computing was in its infancy. New areas of theoretical investigation have developed, leading to methodologies that can be implemented only because of the new computing power. Approaches to statistical inference are in a process of change. In many contexts, Bayesian methods have been too computationally demanding to consider. This is now slowly changing, with the development of software that makes it easier to apply Bayesian methods.

3.3. Connections into Other Software

Some of the largest changes that are in progress have to do with connections into other software programs, and in connections with other systems and especially with the internet. A major concern of developers of statistical computing systems is to connect statistical software into other software that will be used in the course of a project, in a way that will simplify the total task. Tasks that are more complicated than necessary may not be done well, if they are done at all. Some tasks, now thought highly necessary, would be almost impossible without the assistance that computing systems give.

The demand for connectivity goes in both directions. Not only do users of statistical programs wish to be able, easily and seamlessly to link into, for example, database systems. Many users of database systems have a demand to link straightforwardly into software that has statistical abilities. This is much better than the inclusion of inevitably limited and selective statistical abilities in the database system, done without access to high levels of professional statistical scrutiny. Database systems are by no means the only such example. Here are some major types of program that, in 2004, it would often be helpful to connect into statistical software systems:

- Data handling systems, i.e., database systems;
- Document processing, display and search systems; i.e., text editors, word processors, text markup systems, graphics systems and display systems;
- Project management systems;
- Human interface systems;
- Networking and internet connection systems.

Database technology, which has grown up alongside statistical system technology, is now a relatively mature technology that has widely implemented evolving standards. Just as has happened with statistical systems, new areas of scientific activity continue to create new demands. There has long been an acceptance of the importance of interfacing statistical systems with database systems.

-
-
-

Bibliography

Baldi, P. and Brunak, S. (2001). Bioinformatics. The Machine Learning Approach. MIT Press. [This book canvasses a range of statistical and computational methods that are important in computational biology. In this context "machine learning" is largely co-extensive with statistics. A weakness of this book is its very limited discussion of model criticism.]

Dominici F., McDermott A., Zeger S.L. and Samet J.M. (2002). On the Use of Generalized Additive Models in Time Series of Air Pollution and Health. *American Journal of Epidemiology* 156: 193-203. [This reports on an ongoing series of studies that used Generalized Additive Models to relate mortality to air pollution, with major implications for public health decision making. Unfortunately, the default convergence criteria used in fitting these models were insufficiently stringent. All calculations had to be redone, leading to substantial corrections to estimates of the effects of small particles.]

Ewens, W.J. and Grant, G.R. (2001). Statistical Methods in Bioinformatics. Springer. [As the title suggests, this is a wide-ranging account of the application of probabilistic and statistical methods in bioinformatics.]

Gentleman, R. and Carey, V. (2002). Bioconductor. Open source bioinformatics using R. *R News* 2: 11-17. http://cran.R-project.org/doc/Rnews [Describes the open source Bioconductor project, intended for use with microarray and other genomic data.]

Gigerenzer, G. (2002). *Reckoning with Risk. Learning to Live with Uncertainty.* Penguin Books. [Discusses the need for clear thinking in connection with important public health issues – screening for breast cancer, AIDS testing, and so on. This is an excellent introduction to simple uses of Bayesian methodology, in contexts where such methodology is essential.]

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Krüger, L. (1989). *The Empire of Chance*. Cambridge University Press, Cambridge, U.K. [This is an interesting and well-written history of the development of statistical ideas, describing the different historical origins of different approaches.

Gower, J.C., Simpson, H.R. and Martin, A.H. (1967). A statistical programming language. Applied Statistics 16, 87-89. [Describes early work in the construction of a language for handling statistical data manipulation and statistical calculation.]

Gradstein, F. M., Ogg, J.G., and Smith, A. G. (2004). A Geologic Time Scale 2004. [Describes a major revision of the geologic time scale, with details of the statistical methods used.]

Higham, N.J. (1996). Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia. [This is a well-written text on numerical computation.]

Lang, D. T. (2000). The Omegahat environment: new possibilities for statistical computing. *Journal of Computational and Graphical Statistics* 9: 423-451. (See also http://www.omegahat.org) [Describes the Omegahat project, which is intended to be an umbrella for the development of a new generation of statistical systems and software.]

Levenez, E. (2003). Computer Languages Web Page. http://www.levenez.com/lang [This is a useful and comprehensive resource for information about statistical languages.]

Maindonald J H (1992). Statistical design, analysis and presentation issues. *New Zealand Journal of Agricultural Research* 35: 121-141. [Discusses common problems with the way that statistical methods are used and results presented, and sets out principles that should guide researchers in their use of statistical methodology.]

Ripley, B.D. and Fei Chen, R. (2003). Data mining by scaling up open source software. Invited paper, Proceedings of the 2003 session of the International Statistical Institute. [Discusses issues that arise in interfacing a statistical system to large databases, where it may be important to keep to a minimum traffic across an internet connection.]

Selfridge, P. (1996). In from the start. IEEE Expert 11: 15-17 & 84-86. [Selfridge reflects on the hype that accompanied early work on artificial intelligence.]

Schatzkin, A., Kipnis, V., Carroll, R. J., Midthune, D., Subar, A. F., Bingham, S., Scholler, D.A., Bingham, S., Troiano, R. and Freedman, L. (2003). A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based observing protein and energy nutrition (open) study. *International Journal of Epidemiology*, 32:1054–1062. [Compares results from the Food Frequency Questionnaire (FFQ) with a 24-hour recall method and with the use of highly accurate biomarker measurements. The study has in mind the use of the FFQ, perhaps supplemented with occasional 24-hour recall data, in diet-disease association studies.]

Wilkinson, L. and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanation. *American Psychologist* 54: 594-604. [This paper offers advice on standards that papers that appear in psychology journals should follow.]

Biographical Sketch

John Maindonald after earlier experience as a schoolteacher and University lecturer, worked for 25 years as an applied statistician in publicly funded science in New Zealand. In 1996 he moved to Australia, first to the University of Newcastle and then to Australian National University. He has had wide experience in the use of statistical methodology in many different application areas, including entomology, horticulture and clinical medicine. He has coauthored numerous papers with application area specialists. His first book - Statistical Computation - was published by Wiley in 1984. He is the senior author of a second book - Data Analysis and Graphics Using R: An Example-Based Approach - that was published by Cambridge University Press in 2003. Currently, he is employed by the Centre for Bioinformation Science at the Australian National University, working on methods for the analysis of microarray and other gene expression data.