Jour. Ind. Soc. Ag. Statistics
57 (Special Volume), 2004 : 11-14

Computer Technology and Statistics

Jagdish S. Rustagi
The Ohio State University, Columbus, Ohio, USA

SUMMARY

The article illustrates the changing scene in statistical practices with the extensive application of latest computer technology. The emphasis is laid down on the application of recently developed computer intensive statistical procedure such as data mining, jackniffing, bootstrap, pattern recognition and image analysis. The role of these have also been pointed out for genomic studies.

Key words : Computer intensive methods, Genomic statistics, Microarrays.

1. Introduction

Statistical sciences respond appropriately to the new demands of research and development in various areas of experimental sciences. Theoretical advances for such demands were made in terms of statistical techniques developed for the purpose. In the early development of statistical science, these advances were heavily dependent on mathematics and they helped in the applications of statistics to a wide variety of problems in the sciences. With the advances in computer technology, many innovative techniques for use in applied statistics have recently been developed. In medicine, agriculture, business and government, these applications have solved important problems.

Here a few examples are given to illustrate the changing scene in statistical practice using computers. The high speed of present day computers has allowed statisticians to demonstrate theoretical results as well as to obtain optimal properties of many procedures. With the easy availability of statistical software of various kinds, statistical procedures have been widely used in scientific research, government and industries.

Recently developed statistical procedures such as those of data mining, jackknifing, bootstrap, pattern recognition and image analysis are essentially computational. These procedures have contributed significantly to the solution of many practical problems. For example, in developing models for lead burdens in studying environmental health effect on humans, the solution of Wiener-Hopf integral equation is central and that was accomplished by utilizing numerical approximations of integral equations, Rustagi [6].

In the study of compartmental models in pharmacokinetics and biopharmaceutics recourse has to made to computational techniques since the models are highly nonlinear, Rustagi [7]. In the case of quality improvements in industry where applications of Taguchi methods have been made extensively, for example, solutions in case of multi response case cannot be obtained without computers, Rustagi *et al.* [8]. In numerical optimization, design of computer experiments, industrial experiments, bias reduction in density estimation and other related areas in modern applications, computer solutions are essential for practical answers, Rustagi *et al.* [9]. In the new emerging areas of genomic statistics, bioinformatics and microarray analysis, it seems very likely that without computers, major problems cannot be solved. For example, the multivariate data generated by microarray experiments is high dimensional having thousands of components.

From telecommunications to manufacturing in industry, computer technology is playing a major role. Computers are being used in film making, music reproduction, art and architecture. Significant changes in the conduct of international commerce are being made as it is being conducted on the internet. World Wide Web has been responsible for information explosion on the internet and search engines like Google have brought vast amount of information to the individual consumer by computer. It has allowed distance learning and has simplified the contact of people with their governments. The internet has stimulated cooperative activities among scientists situated at far corners of the world. In war settings, computers have provided the soldier with important tools such as global positioning systems.

The speed of computers is also increasing at a fast rate. The proposed IBM super computer available in the year 2004 is reported to have 12,600 microprocessor with speed of 100 teraflops (10 trillion calculations per second). The storage capacity of some of their machines available in 2005 is supposed to be equivalent to ten billion books.

2. *Computer Intensive Methods*

The wide applicability of computer intensive methods in statistical practice has made significant progress due to the availability of high speed computing. A brief survey of these methods in agricultural statistics was given in a paper by Rustagi [7]. With the ease of computation, complicated problems can be solved in practice. Many problems in Bayesian statistics, require evaluation of integrals which can be accomplished by numerical evaluation. Software for commonly used procedures are available and have resulted in the use of these procedures. Graphical procedures are also easily implemented with the help of computers. In pattern recognition and bioinformatics, use of graphical procedures is usually made. They also help in communicating complicated numerical results to general public. Visualization of high dimensional data is possible through graphical procedures and the advances in image analysis have depended on high speed computing. Automatic recognition of various aspects of human face, for example, can be accomplished now with computer intensive methods.

3. Genomic Statistics

An important area of scientific study that has a potential of significant advances in the near future with the help of computer technology, is genomic statistics. The entire DNA content of a cell including all the genes and all the intergenic regions, is called Genome. Recent completion of the genomes of various organisms including that of humans has been accomplished. For example, in the human genome, three billion sequences have been determined. The study of genome has created another area of study, known as micro array biotechnology that allows the simultaneous measurement of the expression levels of thousands of genes. This technology allows the measurement of the intensities of genes and helps in classifying as well as elucidating their pathways. New statistical problems have arisen in the study of microarrays. There are many applications of micro array technology, for example one can list significant genes whose expression characterizes each diagnostic class. The area of the study of extremely large data sets of genomic information such as DNA sequences and expression from DNA microarrays has evolved into a new computer discipline, known as bioinformatics. The science of bioinformatics will help in many areas of investigations such as human identification for forensic and parentage issues. It can be applied to detection of genes affecting a particular disease as well as economic traits in plants and animals and varietal protection of plants. In agriculture, bioinformatics is likely to play a central role in the age of computer technology.

Statistics has been described as a key technology of the twenty-first century. It is being utilized with the help of computer technology in security matters, chip design and computer manufacturing. Various problems in genomic statistics have been recently discussed by Liu [5] and Zhang and Shynylevich [10].

Software programs have been made available for analyzing various kinds of microarray data. For example, SAM-Significance Analysis of Microarrays has been made available by Tibshirani, Tusher and Chu and is based on their paper in the Proceedings of the National Academy of Sciences. This software is essentially meant as a tool for data mining of microarrays. Another software from the same authors provides classification and prediction errors via cross validation and gives significant genes. This software is called PAM-Prediction and Analysis of Microarrays. Many other programs for computations are also available. Some selected references are given at the end.

REFERENCES

[1] Brown, P.O. and Bostein, D. (1999). Exploring the new world of genome with DNA microarrays. *Nature Genetics*, **21,** Suppl., 33-37.

[2] Kendziorski, C. M., Zhang, Y., Lan, H. and Attie, A.D. (2003). The efficiency of pooling mRNA in microarray experiments. *Biostat.*, **4**, 465-477.

[3] Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819-837.

[4] Kerr, M.K. and Churchill, G.A. (2001). Experimental design for gene, expression microarrays. *Biostat.*, **2,** 183-201.

[5] Liu, Ben-Hue (1997). *Statistical Genomics: Linkage, Mapping and QTL Analysis.* CRC Press, Boca Raton, Florida.

[6] Rustagi, J.S. (1966). An application of queueing theory to a problem of environmental health. *Metrika*, 137-144.

[7] Rustagi, J.S. (1990). Computer intensive methods in agricultural statistics. *Jour. Ind. Soc. Agril. Stat.*, 256-276.

[8] Rustagi, J.S., Singh, V.P., Ghosh, S.P. and Wong, Kwang (1993). *Multivariate Generalization of Taguchi's Signal - to Noise Ratio for Industrial Experiments, Quality Through Engineering Design.* Way Kuo (Editor), Eisvier, New York, 181-190.

[9] Rustagi, J.S., Javier, W.R. and Victoria, J.S. (1991). Trimmed Jackknife - Kernel estimate for the probability density function. *Comput. Statist. Data Anal.*, **12**, 19-26.

[10] Zhang, W. and Shumulevich, I. (Editors) (2002). *Computational and Statistical Applications to Genomics.* Kluwer Academic Publishers, New York.