TRENDS IN STATISTICAL COMPUTING*

S.M. ALVIAR**

INTRODUCTION

For almost half a century, statistical analysis is done on desk calculating machines. During the fifties there were some improvements from hand operated to electric driven calculators. The depth and breadth of the application of statistics during that period were conditioned by what these machines can do. As a consequence, the theoretical advances were held back, not by the shortage of ideas or of capable statisticians who can explore new frontiers, but by the technological impossibility of performing the necessary computations. The advent of computers in the early sixties, opened the floodgate for the advancement of both the theoretical and applied statistics. However, this advancement was also accompanied by problems worth the attention of statisticians and other professionals.

This paper presents the evolution of statistical practice at UPLB. In general terms, the change in the practice brought about by the introduction of modern computing facilities is discussed. This includes the effect on the quantity and quality of statistical practice and the danger it poses on the prevalence of abuse and misuse of statistics. To provide bases in resolving the issues, a survey of approaches is presented. This requires cooperative effort between statisticians and computer specialists and the application of the current trend in computer technology.

STATISTICAL CONSULTING SCENES AT UPLB

In the early sixties, researchers consulted statisticians on practically personalized basis. The first encounter would deal with the detailed discussion on the nature of the research problem from which the statistician based the recommendation for the appropriate experimental or survey design (Figure 1). The sampling frame and the questionnaire, in casc of sampling surveys, and the field layout and data recording table in case of experiments were among the requirements being readied. During the same consultation period the statistical analyses and the methods to be used were also prescribed. During field operations, no further contact with the statistician took place not unless unforseen problems within the domain of his expertise arose, e.g. what to do about lost or damaged experimental units or missing survey respondents.

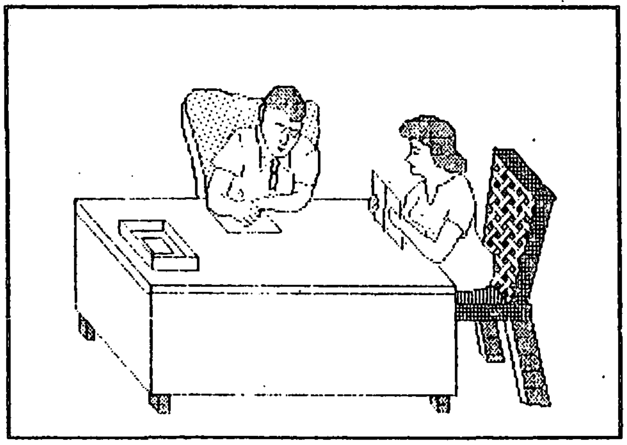


Figure 1. Communication between the researcher and the statistician ensures high level quality of statistical practice.

* Paper presented at the National Conference of the Philippine Statistical Association at Forestry Auditorium. UPLB, College, Laguna on August 14, 1987.

** Associate Professor in Statistics and Computing Science, Institute of Mathematical Sciences and Physics, College of Arts and Sciences, UPLB.

The next encounter by the researcher with the statistician was during the processing of the gathered data. At that time only the Statistical Consulting and Computing Service Center of UPLB has the computing facilities. The more knowledgeable researchers on statistical analyses relied on the services of this office. This resulted to a very close overseeing of the whole affair pertaining to the use of statistics at the University. With this arrangement, the office was able to gather relevant data useful in designing future experiments and surveys. Furthermore, the statisticians were given deeper insights on the nature of both applied and theoretical problems of their research endeavor.

The above arrangement, however, was shortlived as the number of researchers, students and scientists alike, has increased dramatically. The Center's resources were outgrown by the increasing demand that it has to relinguish some of its functions. For instance, the consultation on design and analysis of experiments or surveys was made optional. To offset possible deterioration of the quality of statistical practice, the remedial step taken by the University was to incorporate statistical courses in the curriculum of students and to offer short training courses for researchers. The latter became a regular activity of the office which resulted to the widespread of statistical practice particularly among the universities and colleges in the countryside where UPLB has cooperative projects. Considering, however, that the training course usually covers only the basic concepts and very little time spent on theories and applications, misuse and abuse of the statistical practice were expected to be prevalent among irresponsible users. Thus, it is not uncommon to find researchers seeking the assistance of the statistician when the problem is already beyond resuscitation.

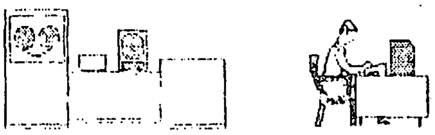
The installation of an IBM S/1620 Model II in the late sixties introduced a new dimension in the practice of statistics at UPLB. The more obvious was the apparent liberated athmosphere in the use of more sophisticated and modern statistical methodologies. Initially, practitioners became visibly communicating with statisticians and computer specialists in their quest for competence in the use of statistical software. Understandably, once a certain level of knowhow has been achieved these practitioners ceased to consult both specialists and do their own data-analytic computing themselves. As a result the problem of control over the unwise practice of statistics subsisted and became more extensive. The proliferation of statistical software of unknown quality further aggravated the situation.

Over the years, this statistical software has germinated and grown without control or management, Fransis [2]. Little has been known of the number of these programs, their detailed characteristics, and which ones are more dangerous. Some developers gave insufficient attention to accuracy and to methods of protecting the user against his misusing the program. User's on the other hand, in publishing the results of analyses, typically fail to identify precisely the program and computer used, Francis commented further.

During the seventies, another major breakthrough in data-analytic computing was the consolidation of isolated programs into integrated programs known as statistical computer packages. These statistical packages provide common formats for command languages and data entry. They are capable of easy data validation and file manipulation. They are far more versatile, easy to learn and easy to use contributing greatly to increased productivity. However, this state of affair tolerated one to engage in the analysis even without proper statistical and computational knowhow which is far more dangerous than the earlier practice.

FROM VACUUM TUBES TO SUPERMICROCHIPS

By way of microminiaturization, what used to be a roomful size is reduced to a notebook or lap-held size computer. The first electronic computer, ENIAC, costs $US5 million to build (about $25 million in 1987 currency), occupied a space of 90 cubic meters, (9m x 5m x 2m), weighed 30 tons, used 18,000 vacuum tubes - of which some hundreds had to be replaced each day - consumed as much power as a locomotive (about 140 kilowatts, a month consumption of an average size family), and generated considerable heat. A modern microcomputer costs $2500 (a clone in the Philippines costs about $1000), is 3000 times smaller, 10000 times cheaper, 5000 times lighter, with a mean downtime (failures) measured in years, uses 2600 times less power, and generates very little heat. The microcomputer is forty times faster than ENIAC, and its memory capacity is 400 times greater. Please see Figure 2.



	ENIAC	MICROCOMPUTER	FACTOR
COST (million $)	25	$0.025	10000
SIZE (cu. m.)	90	0.03	3000
POWER (kw)	140	0.05	.2800
WEIGHT (tons)	30	0.006	5000
MEMORY (kbytes)	2.5	1000	1/400
SPEED (mips)	0.01	0.4	1/40
COMPONENT	vacuum tube	microchip	
DOWNTIME	days	years	

Figure 2. Comparison between the first (ENIAC) and the fourth generation (microcomputer) computers.

Today's faster machines are capable of 24 mflops or millions of floating point operations per second. It is evident that micro- electronic advances permit an exponential rise in output with an exponential fall in inputs energy, labor capital, space and time.

Nothing can parallel this economic phenomenon according to Nora, [6]. To expect a greater return from a smaller investment in time, money and effort seems contrary to every conventional precept. If we dig a trench with a teaspoon, the effort required for each quantity of soil removed would be very small but the time taken to complete the task would seem endless. On the other hand, if we use a spade, far more effort is put into every movement but the task is achieved more quickly. But imagine that we could dig the trench with no more effort than to wield a teaspoon but in far less time than would be taken with a spade.

As statisticians, Kendall, [3], viewed the future of statistics along with the power of computers. He noted, however, that statisticians and practitioners of the art have been too slow to take advantage of the computer. The reason for which is not the limited access to machines but is partly due to conservatism in habits of thoughts. He hoped that this attitude is dying out of its own accord. If not, he suggested that effort should be exerted to kill it.

There are other important points emerging from the new order of the affairs of professionals. First, today's nature and magnitude of problems being undertaken require blending of skills from many disciplines, over which no single expert can have a full command. As an aftermath, it was observed that aggroupation of experts and specialists dominate the realm of institutionalized professional services. They work as a team to assure greater success in solving both theoretical and practical problems.

The same observation is evident among more aggressive academic institutions in assigning research projects to students. Nowadays, students are grouped as a team to work on research projects of wider scope which require deeper intellectual capacity with the end in view of achieving far reaching excellence and relevance in their undertaking. In the old order, where students work independently, areas of focus did not only proliferate but also prevented deeper penetration into the hearts of the problem. As a consequence, it was quite common to find results of their efforts accumulating dust in the archives of the library.

As may have been dictated by the ever increasing sophistication of the new world order, there emerge new subject matter fields in practically all areas of human endeavor. This move was further motivated by the desire to bring closer to human affairs the methods of sciences. Related to statistics, these new fields are technometrics, psychometrics, biometrics, econometrics, sociometrics, etc, to name a few. It is interesting to note, however, that all these fields are branches of scientific methods or are concerned with the application of scientific method in specific fields. Since statistics is embraced in scientific method, in that sense this emphasizes the enormous scope and extent it plays. These expanding horizons of application dictate that we have to sharpen further the existing statistical tools, invent new ones, and learn to use to full power what the computer can offer. Again Kendall said, "It is as though we had gone in twenty years from the spade to the bulldozer. But the operation of diggings remain basically the same". The implication is that we really have to harnessed to the fullest the statistical computing power provided by modern computers to elevate to greater heights our throughput capabilities.

FROM DATA PROCESSING TO EXPERT SYSTEMS

Data processing refers to the use of computer for processes which do not require human interaction during implementation. The steps involve are fixed and routinary, e.g. payroll and accounting reports. Computing, on the other hand, requires human interaction with the system as dictated by the initial results of the process. This is typical of statistical computing where the final output may not be obtained in a single run. Note that data processing suits well the old batch mode operating system while the latter is more appropriate in an interactive mode or personal computer.

Information processing has been elevated to knowledge processing which in turn gave rise to expert system. An expert system is a computer program that has a knowledge of a specific area of expertise which are used to solve problems at high level of performance, similar to that of a human expert. Expert systems are also called knowledge based system since their performance depends on combining principles of artificial intelligence (AI) programming with specific knowledge domain. The scheme is illustrated roughly in Figure 3.

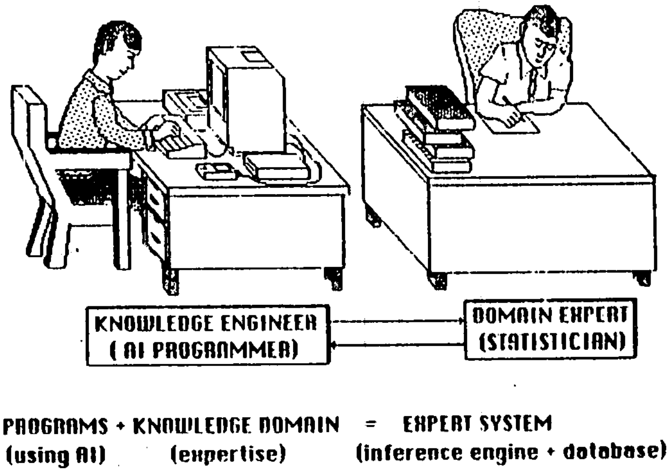

KNOWLEDGE ENGINEER
(AI PROGRAMMER)
DOMAIN EXPERT
(STATISTICIAN)
PROGRAMS + KNOWLEDGE DOMAIN = EXPERT SYSTEM
(using AI) (expertise) (inference engine + database)

Figure 3. Components of statistical expert system.

Expert system uses a fundamentally different method in solving problems. For complex ones requiring experience, the potential solution space can be quite large that makes it impractical to store or search for a problem solution using a conventional software. The approach taken by expert system is to accumulate a base of necessary facts and inference rules. By applying the inference rules to the facts to develop new facts or chaining them together to simulate how one piece of information leads to another, the expert system generates the necessary solution.

The knowledge embedded in an expert system must already exist before system development. Thus expert systems are only written to solve tasks that people are currently able to solve. An expert system is developed through repeated interactions of knowledge engineer and a domain expert. The domain expert is a specialist in the field to which the expert system is applied. The knowledge engineer or expert system designer is responsible for eliciting from the domain expert and quantifying how he solves the problem and then formalizing it using AI programming techniques. The knowledge acquisition is usually a tedious process requiring technical and diplomatic skills of both the knowledge engineer and the domain expert. Knowledge acquisition is currently the bottleneck in building expert systems. It should be mentioned that some of the goals of expert systems are not only to use knowledge effectively but also to encode it in active form.

In the synthesis of COMPSTAT 86, Marbach, [4], pointed out that statistical expert systems are new developments, still in an early phase, but the initial steps have been taken and more effort should be exerted for it to progress. In the same report, Muxworthy said that expert system will come but major breakthroughs are needed before it can be realized. However, one contributor argued eloquently that resources would be much better spent at the present time on teaching young statisticians rather than an expert systems development.

Gruger and Ostermann, [5], claimed that constructing a statistical expert system for a small, well defined statistical problem, where knowledge consist of a small number of hard decision rules is no problem. Such systems are equally well suited for the statistical expert and the statistical novice. They strongly suggested that emphasis should be given to the graphical component of expert system. This, according to them, will provide a means of offering the user a reliable information about the data making it easier to supply further information about the underlying situation which the expert system may fail to detect on automatic basis. For example, Figure 4 illustrates the convenience of assessing the quality of fit of the quadratic model surface on the observed suspended points.

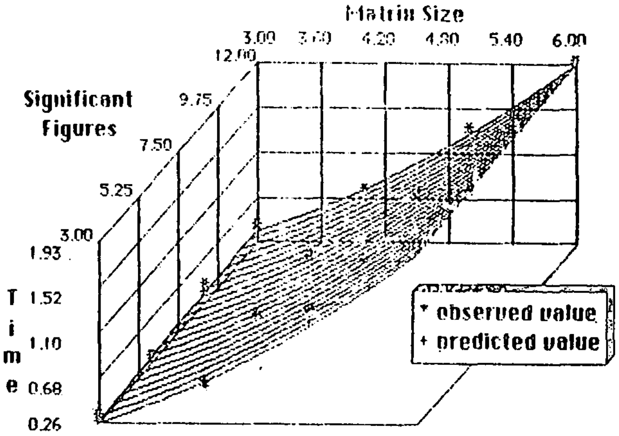


Figure 4. The effect on execution time of the matrix size and number of significant figures for Gauss-Jordan inversion algorithm

Caringal and Albacea, [6], developed an expert system for experimental designs aimed at making available the knowledge of an expert in the field of experimental designs to novice researchers. The systems not only assists in designing experiments but also provides the facilities for analyzing the collected data. While it is still far from being considered as a full pledge expert system, a continuing effort towards achieving this end can be programmed.

SUMMARY AND RECOMMENDATION

The quantum leap in the progress of computer technology advanced correspondingly the frontier of statistical theories and applications, along with some attendant problems. A case in point is the deterioration of the quality of statistical practices by non-expert users brought about by the availability of easy to use and easy to learn statistical computer packages of unknown quality. With easy access to computers or economically affordable microcomputers, the state of affair tolerates liberal application of statistics even by the uninformed and unguided practitioners thus proliferating misuse and abuse of the practice.

On the other hand, computer science provides new tools for statisticians and clients which are more than the mere extension of previous capabilities. In the area of databases, integration of database management and statistical analysis system is a big step towards development of more comprehensive software packages with semi-expert systems capabilities.

Expert systems have been the subject of interest in many areas of scientific disciplines, apart from the computer science. With the use of this new technology, the problem of statistical malpractices can be alleviated by elevating the capabilities of statistical software packages. This can be done by imbedding into these softwares the functions and expertise served by professional statisticians. It is imperative that there is a need for closer cooperation between statisticians and computer specialists as a prerequisite to success in this undertaking.

REFERENCES

[1]. Chambers, John M. (1981). Some Thoughts on Expert System. Computer Science and Statistics: Proceeding of the 13th Symposium on the Interface, Springer - Verlag, New York.

[2]. Francis, I. (1981). Statistical Software a Comprehensive Review. Elsevier North Holland Inc., New York.

[3]. Kendal, M. G. (1972). The History and Future of Statistics. Statistical Papers in Honor of G. W. Snedecor, Ed. T. A. Bancroft, Iowa State University Press

[4]. Sint, P., Marbach, G., Muxworthy, D. and Y. Escoupier (1986). A Report on COMPSTAT 86. Ed. A. J. Westlake, Statistical Software Newsletter, Vol. 12, No. 3, 107 - 110.

[5]. Gruger, J. and R. Ostermann (1986). Construction and Integration of Statistical Expert System for Binomial Experiments. Statistical Software Newsletter, Vol 12, No. 3, 124 - 128.

[6]. Caringal, R. B. and E. A. Albacea (1987). NANO: An Expert System for Experimental Designs, Paper presented at the Fourth National Convention in Statistics, June 15, 1987.