STATISTICS, THE COMPUTER AND OUR MODERN SOCIETY

BY

CARL F. KOSSACK

F.A.O. Consultant at I.A.R.S.

On several occasions I have had the opportunity to speak to professional groups relative to the impact that the modern high speed computer will have upon their profession but this is the first time that I have attempted to address such a consideration to a group within my own profession of statistics. However, I believe that such a consideration is more than appropriate and I have not only been giving this problem considerable thought but have entered into many discussions about the problem with my statistical friends and colleagues. I hope that my views on the problem will stimulate thought and discussion here in India.

When I introduce the computer to an audience, particularly if it is a lay audience, I stress the revolutionary effect the computer has already had upon the areas of numerical computation and data processing; noting that, the speed of such activities has in less than two decades increased by over six order of magnitudes. With some audience, I recall with them that this means that ten to the sixth power is used as a divisor or that the time required for such operations is one millionth of what was required less than 20 years ago. In fact, it is interesting to note that only in the field of energy conversion through the developments that have taken place in the nuclear energy field has any other such startling a change in capability taken place. In the field of transportation, for example, though the modern jet aircraft has made the world smaller, the development of air flight during the last 60 years has not increased our transportation capability by even two orders of magnitude. In

*Tehnical address delivered at the inaugural session of the 23rd annual conference of the Indian Society of Agricultural Statistics held at Bombay on 22nd December, 1969.

computing and in nuclear energy we find however this revolutionary force impacting our society and one soon apprrciates that in both of these areas man is faced with challenges that must be met if our society, as we know it, is to survive and prosper.

Still another aspect of our present-day situation needs to be recognized if we are to obtain a proper view of the challenge that now faces statistics or persons that think of themselves as statisticians. That is the increasing industrialization and urbanization of our society with the resulting need for a better and faster response to problems that such large population and production concentrations create. The dynamics of our modern society demand rapid but meaningful decisions even in the fields of research and these always must be made in the face of uncertainity. Herein lies the challenge to statistics. Can we as statisticians harness the speed and power of the modern computer in such a way so as to be able to respond to this challenge of modern society ?

I do not know whether it is at all necessary for me to document the above considerations, but let me, for the sake of discussion, briefly look at the field of medical research. There are any number of examples of how our modern society has insisted upon having research findings flow rapidly into medical practice. Consider the situation relative to the development of new drugs or vaccines. Testing programs now are required to be of short duration so that the new product can be placed on the market at an early date either because there is such a competitive advantage to be gained through such action in a free economy system or because there is the need to show early success in a government sponsored program. Even in surgery we note that the recent advances in heart surgery when introduced in human medictine brought considerable professional complaint of too much haste but the public seemed to applaud the efforts. It is apparent that if such a need for shortening the response time exists in the research field that in the more industrial and production oriented activities, the need to shorten the response time has now become critical. Thus we find, in electronic manufacturing that inventory decision evaluations are required on a daily or even more frequent basis whereas a few years ago an annual inventory sufficed.

If statisticians are to continue to accept the role of providing scientific based procedures to decision makers, it seems appropriate to critically examine the question as to whether statisticians are

facing up to these new rapid response time requirements that modern society is placing upon its decision makers. Before considering this question, I believe we should bring into focus the added capabilities that the new third generation computors have given man to use in his computations and data processing. Two such capabilities can be mentioned here, one is the introduction of significantly larger and more rapidly accessible memory system in the form of massive data banks and the other is the development of remote terminal systems with display capabilities involving man-machine conversational implications. Both of these new capabilities will have a major impact upon statistics since with the availability of really massive memory banks it is to be expected that a continuing type approach to information file maintenance will now become feasible and with a conversational mode being possible through the use of the new remote consoles, a researcher may be expected to evolve his analysis in a more sequential fashion and thus he will want to develop models that more closely match not only the actual conditions underlying the problem but also those that are exhibited by the observed data.

With this as background, we can ask how effectively has the statistician adopted to the computer as a tool of analysis. If I were forced to give an answer to this question based upon my American experiences, I am afraid that I would have to reluctantly admit that up to the present the statistician has not really faced up to the challenge of the new computer age. May I support this evaluation with several generalities before getting down to more detailed considerations.

First, we can note that very few major computer centers in America are organized as part of a statistical laboratory, and even those centers that initially come into being as part of a statistically oriented program now have separated off to become an independent unit often even with a less tangible linkage to statistics than that found in other centers. In Universities this phenomena is so common that I will not even mention any examples, while in research laboratories and in industry the role of the statistician is often not even recognised sufficiently to warrant a separate organization structure for statistics, let alone to have statisticians involved in the activities of the computer.

In addition, we find that the statistical subroutines that are available within computer centers are at best of a "canned" type usually involving a static model for the analysis and more often

than not such routines have not really been completely checked out by the computer center personnel. Thus, we find that if a researcher wants a statistical analysis performed through a computer center he will be gixen a standerdized analysis often with pages and pages of output which rarely is appropriate and often not even appreciated by the researcher.

Finally, the quality of statistical staff associated with computer centers, if in fact there is such a staff, is usually of a immature and poorly trained type since the professional statisticians of an organisation do not concern themselves with such mundane details as computing. For example, at a recent international conference involving a special area of statistical analysis, although most papers presented included a section on numerical evaluations which had been processed by a computer, only one or two of the speakers had even the faintest idea of how the computer analyses were actually accomplisoed and most speakers were lucky that they could read the output summaries correctly.

Thus we have the situation that a dramatic and powerful analytical tool is impacting every branch of our science and technology and the statisticians have yet to respond to the challenge. I feel like I should emulate our American patriot "Paul Revere" and get on a horse and ride through the streets of our statistical community shouting "the computers are coming". However, they are already here.

To illustrate what type of adjustment is needed in statistics let me consider three different areas of statictics within which one can easily make changes and improvements now that the computer is here to stay. The first area I will call data processing and parametric modelling since most of the applications that will be mentioned are of this nature. Some time ago, I reviewed the set of available statistical computer routines to determine what advanced applications might be available to individuals through the use of such general computer programs. The result of this effort yielded some seven advanced applications which I would like to first list and then to briefly describe. The list involves :

1. Cluster Analysis.
2. Power Spectrum Analysis.
3. Factor Analysis.
4. Simultaneous Regression.
5. General liner model.
6. Response Surface Analysis.
7. Multivariate Statistical Classification.

Techniques of Cluster Analysis generally provide a recognition of the power of the modern computer in performing data analysis in as much as they require a large amount of non-structured data involving many observations each of which involves a large number of variables. Since so many of our present day data files meet these requirements one would expect that such a technique would be receiving more than its share of attention, but up to now the use of computer programs for making a cluster analysis has been very modest. Essentially what such analyses do is to look for homogeneous clusterings of observed points in the multivariate sample space thus indicating to the analyst which observations "look alike" relative to the variables he has used in his data acquisition program. It is my opinion that such a basic approach to the interpretation of one's data provides a means of giving a preliminary structure to a research problem area. One of the most interesting application of this technique which has brought about a whole new field of research is found in biology and has come to be called "Numerical Taxonomy."

Since, if one were to examine the population of data presently being acquired by our modern society he finds a large segment of these data consisting of analogue data, the use of some type of analysis which would enable one to parameterize such signals is indicated. The use of a power spectrum computer program would provide the basis for such a parameterization but if the digital computor is to be used there must be a digitizing interface between the analogue signal and the computer. The mathematical model underlying the power spectrum analysis assumes that the complex signal is a composite of cosine functions of varying amplitude and frequency. The function of the analysis is to decompose the signal into its principal trigonometric components. In a power spectrum analysis the square of the amplitude is being measured as representative of the power of the particular cosine function in the composite output. Recent advances have introduced a multivariate type model which enables one, to include in his analyses the correlations that exists between two or more simultaneously acquired signals. One need only to note the instrumentation that is presently being utilized in the space programs to recognize the significant role that analogue data is playing in our modern technology.

I have included Factor Analysis within my list of advanced applications though this type of Analysis has been used by psychometricians for almost half a century. The reason for including this type of data analysis is that not only does it gives emphasis to the

need to improve the variable selection mechanism used in most statistical applications but it also illustrates the contribution that the modern computer can make to such applications. In my professional life-time, I have seen the value of a factor analysis reduced from that of providing the essential input for a doctorate degree to that of costing less than a dollar to have one run on a high speed computer. To-day the problem for statisticians is to be able to take advantage of this increasing capability and to use such variable selection routines as factor analysis so as to be able to direct data acquisition energies into more fruitful channels.

Simultaneous Regression techniques enable man to utilize a more meaningful model in his attempt to evolve prediction formulae. In many systems the inter-relationship that exists between the set of observed variables is such that it is difficult to designate which variable really plays the role of the dependent variable. As a result a single regression equation may fail to properly model the system and as a result the coefficients obtained for the single regression model may be spurious. Although the econometricians have long recognised this phenomenon, it has only been recently that valid estimation procedures could be applied when simultaneously dealing with sets of regression equations in which the role of variables may shift from one equation to another. Since similar rogression models can be useful in biology medicine, sociology and political science and in fact in almost all fields of endeavour and since computer programs for making valid coefficient estimates are now available, the challange to researchers in all branches of science and technology is to come to a better appreciation of the interacting mechanisms that regulate system type considerations so as to be able to evolve more appropriate regression models.

The theory now available through the approach generally doneted as the "General Linear Model" in which analysis of variance, analysis of covariance and regression techniques are all incorporated into one analytical model is today providing a challenge to computer programmers and numerical analysts. The numerical problem is concerned with numerical operations involving sparce matrices, since though the model through its comprehensive nature, enables one to handle all of these diverse types of analysis, it generates matrices, with an excessive number of zero elements. Thus the efficiency of the resulting computer routines is quite low and therefore of limited attractiveness. However, once an efficient procedure has been evolved, there will be

available a single computer routine which can handle all types of experimental designs and analyses.

In recent years a new approach to experimental designs has been developed that discards the classical concepts behind hypothesis testing and considers instead how one might best operate a production system so as to obtain optimum results. Since the analysis of these new type designs involves computations such as the solution of sets of simultaneous equations, the use of a computer program for the analyses of these designs is almost mandatory if the dimension of the sample space is large. This type of analysis however indicates the directions that statistics should move if we are to recognize the impact that computers are having in our society. That is more and more attention should be directed towards decision theory as they relate to operational systems. Such a movement will create a demand for statistical techniques and procedures that relate to the estimation procedures and the decisions techniques found in designing and using operational type systems.

The final advanced application that I will mention are those falling within tne area called Multivariate Statistical Classification Techniques. Such techniques evolve a decision rule that will classify an individual into one of several populations based upon the value of a vector of observed variables for the individual. The decision rule becomes of a statistical nature if it depends upon sample data obtained from each of the several populations. Although there have been several different types of classification model available in the statistical literature, these statistical techniques have yet to be effectively used by individuals who are required to make such classifications in practice. This is true even though one can defend the thesis that a majority of man's decisions are of a classificatory nature. Recently the author have evolved a more generalised type of classification technique which enables one to utilize all types of observable random variables in the development of a classification decision rule. A general computer program now has been written for this technique which simply accepts the sample data and then the computer programs itself so as to be able, when called upon in the future, to make a classification decision for one or more new individuals. The problem being studied now involves how to introduce this computerized statistical technique into extensive practical application. Some progress has been achieved in the area of medical diagnosis and in the personnel utilization field.

Let us in fact now turn our attention to this last application in more detail since it involves more of an operation decision rule than simply a data processing for parametric type procedure. It may be fruitful to consider why it is that though man has frequent need for a classification type decision that the techniques evolved by statistician though they have the characteristics of minimizing the probabilities of making a misclassification type error, still are not used. As I consider this problem certain explanations can be advanced that help to understand this paradox. Let us briefly examine some of these:

(1) The models used by statisticians are not in tune with the underlying situations encountered by individual in the real world. In particular the classification models generally assumed that the observed variables are of a single kind, say all continuous or all categorical while the vector of variables observed in real situations are almost always of a mixed nature often containing even analogue type data. Thus the applied person cannot really find an appropriate statistical model to use in his application.

(2) The assumptions made relative to the nature of the distribution of the variables are too restrictive. We so frequently find that in the field of multivariate statistical analysis that one assumes not only that the observations have a multivariate normal distribution but that the several variance covariance matrices are in fact all equal. Though by taking advantage of this assumption, the theoretical statistician can evolve simple models with attaractive estimation procedures whose sampling distributions lend themselves to statistical analysis. The applied person rarely encounters data that are so obliging as to meet such stringent specifications. Now if we couple with this the fact that no general procedure exists for measuring robustness of such techniques it is little wonder that the world is not rushing to apply statistical techniques when real problems are under consideration.

(3) The available statistical data from which the estimates needed to apply the technique must be evolved are incomplete, confused and unreliable. In fact there are so few adequate data files in the world today which can be used in a comprehensive statistical analysis, that it is little wonder that no really significant system type applications have as yet been accomplished. For example not too long ago a large corporation which has kept its operational records over the last ten years on punch cards discovered that it was

impossible to use them in any comprehensive statistical analysis since the changes in codes and the overpunching of fields had completely confused the meaning of the punched records.

(4) Individuals when called upon to provide cost or worth factors to be used in statistical decision techniques are unable to supply such required inputs. Although man with the usual lack of the systematic aposteriori consideration of his subjectively arrived at decisions has no difficulty in making and even stoutly defending such decisions, when it becomes apparent that the utilization of a well structure statistical decision rule is to be used, finds it at best uncomfortable to place values or costs on the possible errors that may be made. This inability clearly exhibits the fact that he has not given such necessary system characteristics any real consideration.

(5) Many organisations, especially at the decision making levels, have insufficient technical ability to apply advanced statistical techniques such as those found in multivariate statistical classification. Also there is a considerable lack of ability to appreciate such consideration. Why is it that although many students receiving training in statistics, few are found in the upper levels of administration ? The rollary result to this deficiency is that few top management groups have available to them statistical talents. One answer to this question may be found in the type of training that professional statisticians are given. Emphasis is given to mathematics, experimental design and related hypothesis testing and distribution theories. Data analysis is rarely included in a formal course, nor do students of statistics receive much business training. In fact in America, there are so few sampling and survey training programs that one is tempted to conclude that this aspect of data acquisition is not considered to be of a quality that warrants study.

Let me conclude this section by summarizing the five factors that I believe contributed, not only to the dearth of application of statistical classification techniques, but also result in there being but modest use of statistics by individuals within the "power structure" of our modern society. They are :

(1) Statistical models do not correspond to reality.

(2) Assumption made relative to the distributions of observed variables are too restrictive and there is no method of testing robustness.

(3) Lack of reliable data for estimation purposes.

(4) No appreciation of the worth or cost factors needed for such statistical analyses.

(5) Lack of technical know-how for applying statistical technique and interpreting the outcome.

Since up to now we have emphasized the pessimistic aspects of our present day approach to statistics, I would like to consider a third area of statistics, that of the use of sample surveys for information file generation and in this consideration to briefly indicate how statistical methods when properly linked with modern computer can meet the growing demands of our modern society. The approach I would like to outline is what I have choosen to call "A Perpetual inventory Approach to Information File Management". In this approach it is assumed that the principal purpose of an information file is to provide data for planning purposes although the existence of such files can also provide information for the operational decision maker. I could spend quite some time discussing the inadequacies of our present census approach to information acquisition, but will let such deficiencies be implied and only enumerate some of the requirements that must be met by any program that attempts to provide planning information. These include :

(1) Information is needed on small units or aggregates rather than on population totals.

(2) Information must be current since planning often considers even weekly or monthly changes.

(3) Objective type variables need to be measured requiring skilled and trained staff.

(4) The required analyses of the information files are of a dynamic nature and involve extensive and sophisticated analytical capabilities.

(5) Evaluations often require a system simulation type procedure.

(6) Technical advisers for the design and analysis of information acquisition programs are needed on a continuous basis.

(7) Continuity of financial support of the program must be assured.

We today find ourselves in the position where modern scientific planning is placing demands upon the information files which can

not be met effectively through the classical sampling survey approach used by statisticians in the past. Fortunately, the advent of modern computer technology introduces a capability that provides one with the means of efficiently meeting these new information demands.

Let us examine the concepts behind the Perpetual Inventory Approach to Information Files. Basically, the approach involves the maintenance within the memory of a computer a complete and current information file involving all units of a population. There is no real storage problem involved as far as modern computer systems are concerned since even computers which still utilize magnetic tapes for mass storage can handle the storage needs for most population information systems. The problems of keeping the information file current are resolved through the use of small sample surveys which are taken at least once a year. However, the heart of the approach is the use of an up-dating mechanism that utilizes the concept found in statistical regression to estimate the changes that have occurred in the unsample units of the population. The regression model is determined by using both the knowledge that the analyst has of the role that various descriptive variables play in creating change along with the actual data on changes that have been observed on the sampled units. Thus throgh frequent sampling surveys and the updating of the information for the unsampled units, the entire information file is kept current on an individual unit basis. Though it is recognized that some of the information in the file will be in estimated form, the use of a rotational sampling plan will soon replace each estimated value by an observed value. The principal aspect of the concept is that of orienting the entire system around a computer and to so structure the related computer system that the information and analytical requirements of scientific planning will be satisfied.

If one considers the population of resource points making up an information file, he can generally identify with each point two types of variables; descriptive variables, which often are either slow to change or can be updated through some legal type reporting activity ; and resource variables, which describe the current level of productivity of the point and are subject to frequent or continual change. These latter type variables are the ones that most frequently enter into planning considerations and thus information on their status is needed on a current basis. We can thus formalize our information file maintenance problem by considering a popula-

tion of resource points identified by $i=1, 2, \ldots$ and I associated with each point is a set of descriptive variables, say x_{ij}, $j=1, 2, \ldots, J$. In addition there exists a set of resource variables, say $Y_{ik}(t)$, $k=1, 2, \ldots, K$ where the value of the resource variable varies with time. At any time, t, one is interested in having available estimates of the value of the y's for all resource points and from these estimates can be evolved for each type resource the total of the resource that is available over any subset of points.

Let us consider for convenience that time can be divided into a discrete set of time period, say $t=0, 1, 2, \ldots, t-1, t, t+1, \ldots$ and restrict our interest in the values of y_{ik} to those that exist at one of these discrete points in time. Let us define $\triangle y_{ik}(t)=y_{ik}(t)-y_{ik}(t-1)$ so that

$$y_{ik}(t)=y_{ik}(t-1)+\triangle y_{ik}(t).$$

Assuming that we have available estimates of $y_{ik}(t), t=0, 1, 2, \ldots t-1$, for all points of the information file the updating of the file can be accomplished if we can obtain estimates of $\triangle y_{ik}(t)$, the change in y_{ik}, for all values of k.

Let us further assume that there exists a regression of $\triangle y(t)$ on the set of x's represented by

$$\triangle y(t)=g(\underline{x}, \underline{y}, \underline{\theta})$$

where $\underline{x}$ represents the vector of descriptive variables, $\underline{y}$ represents the vector of resources and $\underline{\theta}$ represents the vector of regression parameters. In some situations the form of the regression equation and the values of the regression parameters for a particular time period will be available from outside considerations. This would occur, for example when a quota has been established for production of a particular commodity. However, it is expected that, both the form of the regression and the estimates of the values for $\underline{\theta}$ must be obtained from actual experience. In this case one can approach the file maintenance problem using a sampling plan similar to that used in the statistical approach associated with sampling on successive occasions. If one used such a sampling plan, during time period t a sample of resource points would be surveyed and actual values of $y_{ik}(t)$ obtained for these points. For the sample one would therefore have observed values of

$$\triangle y_{ik}(t)=y_{ik}(t)-y_{ik}(t-1).$$

One can use these observed values and an assumed form for the regression to obtain estimates of $\underline{\theta}$ say by the methods of least squares using the observed $y_{ik}(t)$'s and their estimated errors.

$$g_{ik}(\underline{x}, \underline{y}, \underline{\theta}) - y_{ik}(t-1).$$

Using the resulting regression equation one can then update the unsampled resource points for $y_{ik}(t)$, the adequacy of estimates based upon these updated values would depend upon the size of the subset of resource points for which a total is required and upon the strength of the correlation that existed between $\triangle y_{ik}(t)$ and vectors $\bar{x}$ and $\bar{y}$.

In the perpetual inventory approach to the information file maintenance problem each resource point can be assigned a weight according to its importance in the ensuing analyses and the probability sampling plan designed so as to reflect these weights. These weights can be proportionately increased from time period to time period for those resource points that remain unsampled so as to make certain that all resource points are included in some sample within any given number of time periods.

Using the above procedure one will keep available in machine readable form an updated record for each resource point of the population and, in fact, the complete record will also provide estimates of the behaviour of each resource point over time. Combining with this procedure the modern computerized techniques of information retrieval, report generation and simulation one would be able to meet the information and computational needs of planners and policy makers in a dynamic and timely fashion.

The three illustrations described above provide sample documentation of the challenge that our modern society and the computer is making of statitsics and we can repeat the question "Will statistics rise to the challenge ?" Since I believe that the needs of our modren society will be met one way or another and that the modren high speed computer will play a vital role in whatever methodology emerges, the need is for statistician to face up this challenge or to run the risk of being phased out of even our existing role in our modern technology. Think about the present trend of emerging new disciplines. We have Computer Science, Information Science, Numerical Methods, Stochastic Processes, Biometrics. Mathematical Biology, Econometrics. Psychometrics System Analysis, Operation

Research and Management Science just to name a few of the new desciplines. Often what is being developed under these labels either is statistical or should be more statistical.

Need I summarize the messages that I hope my illustrations carried ? (1) Advance statistical applications are currently available that do utilize the power of the computer but they are not generally carried as part of the statistical curriculum and thus receive very little usage though they provide a more powerful mechanism of analysis, (2) Statistical techniques that would provide analysts and decision makers with useful tools are un-useable since they fail to meet the underlying conditions present in the real world, and (3) *A* proper appreciation of the power of a computer if used to structure the statistical techniques evolved to solve real problems can significantly enhance the contribution that such techniques make to our modern society.

The problem is to find a mechanism to reverse the trend and to reorient statistics into more meaningful and practical areas. If, however, we continue the approach of having "the blind lead the blind", that is to have the statisticians whose training and background does not encompass the computer continue to train the young statistical students in the same manner in which they were trained and if statistical research programs continue to stress abstract and impractical considerations because of their mathematical merits, the discipline that we have come to know as statistics, will vanish and perhaps it would be a good thing.