

# ROLE OF STATISTICS AND COMPUTERS IN AGRICULTURAL RESEARCH

S.D. Sharma

I.A.S.R.I., Library Avenue, New Delhi-110 012

sdsharma@iasri.res.in

Agricultural research has played a key role in the development of statistical methods. The presence of wide heterogeneity in the experimental material that is often used in agricultural research, led to the application of statistical tools and consequently many refinements and newer developments in statistics followed. The famous statistician, Sir R.A. Fisher and his colleagues at Rothamsted Experiment Station in United Kingdom and elsewhere, while attempting statistical solutions to agricultural problems, led to the development of design of experiments and analysis of variance techniques which are fundamental to the subject of statistics.

Statistics, in fact, provides scientific tools for representative data collection, appropriate analysis and summarization of data and inferential procedures for drawing conclusions in the face of uncertainty. It is indeed true that statistical tools have wide applicability to almost any branch of science dealing with the study of uncertain phenomena involving aggregates. It has also been established that many apparently deterministic processes, on closer scrutiny, turn out to be inherently stochastic in nature. However, in agricultural research, statistics finds some of the very interesting applications which often led to the development of newer statistical techniques or at least a refinement of existing ones. Consequently, the branch of statistics dealing with agricultural research has been recognised as a separate entity in itself, as **agricultural statistics**, in view of the growth of the subject particular to this area.

Agricultural statistics has three core subject areas, namely, sample surveys, design of experiments and biometrical techniques. Sample surveys in agriculture is primarily concerned with estimation procedures for area under different crops, crop yield and crop production, for various crops over different regions of the country. Besides, the estimation of land use statistics, statistics related to input use in crops such as the varieties, seeds, fertilizer, irrigation, insecticides/pesticides, machines/implements/tools, the supply and demand of various inputs, are often collected through sample surveys. The cost of cultivation/production needs to be compiled through detailed survey inquiries so as to understand the farm level efficiency. The information about markets, prices, imports/exports is also becoming important now that India has attained food sufficiency leading to surpluses in certain pockets and the fair competitive trade discipline being enforced under World Trade Organisation regime. The household consumption surveys, are important for assessing food security in the country. The need for obtaining estimates at smaller area level such as the block level and village panchayat level estimates especially for local governance under the Panchayati Raj system is increasingly being felt. The advance estimates of area sown and crop production is also needed for planning marketing strategies in different regions of the country. Often, complete censuses are periodically conducted over longer time intervals such as quinquennial or decennial, so as to create a sampling frame and also obtain certain primary simple information on population counts, age, sex, holding size classes etc. and are followed by annual/short interval sample surveys for detailed enquiries. Every sample survey involves sample

selection scheme and the estimation procedure. The subject of sampling theory deals with these aspects and the logical foundations/basis for such procedures along with the choice of better sampling strategies using auxiliary information where feasible.

Design of experiments deals with the study of methods for comparing treatments/varieties/factors under different experimental situations that are often faced by agricultural research workers. One obtains data from postulated, hypothetical, infinite populations using the basic and essential principles of randomization, replication and local control as enunciated by Sir RA Fisher while laying down the principles of good experimentation techniques. The principle of local control in the form of identifying homogenous blocks (an equivalence of stratification in sample surveys) is a means to control error variation while generating data. This enables the researcher to properly account for the variability in the experimental material. The technique of analysis of variance was developed by Fisher basically in the context of analyzing data from well designed experiments, so as to partition the total variation into components due to assignable causes and those due to non assignable random causes also termed as error variation. The emphasis is on reducing error variation employing local control and if possible, through additional information on concomitant variable, closely related to cause of variation, employing a technique known as analysis of covariance. Besides, the experimenter is also interested in making certain specific comparisons among the treatments tried in the experiment. This is achieved by using the technique of contrast analysis.

Sometimes, the agricultural experimenter may be interested in establishing relationships to explain the behaviour of the variable under study as a function of certain input factors. Even though, the functional form of the behaviour may be unknown but the same can often be approximated by a linear function as a simplifying assumption. Under certain assumptions on error distribution, using the method of least squares, one can always fit the functional relationship, termed as Regression function, providing for the establishment of such relationships, and related issues of estimation and the test of significance of parameters involved.

Although intuitive appeal and the scatter diagram help in selecting the best model, it is still important to check that the chosen model is a reasonable description of the data arising from the experiment. Model assumptions can be checked by studying residual variation, standardized residuals (dividing by their standard deviation) and residual plots. Following checks are often useful:

- Check the form of model – whether the form chosen is appropriate or an alternative form could be better?
- Check for multicollinearity – are the input variables correlated?
- Check for outliers – are there any unusual observations or outliers?
- Check for independence - do the error terms appear to be independent?
- Check for constant variance - do the error terms have the same variance?
- Check for normality – do the error terms appear to be a random sample from a normal population?

After fitting the response model one likes to obtain the levels of input variables at which the response is maximum. Often, the well known methods of maxima and minima are used to obtain global unrestricted maxima (if it exists), for a quadratic response function.

Alternatively, linear programming techniques could be used for obtaining an optimum with respect to specified set of linear constraints.

Lastly, the biometrics is another core area under agricultural statistics. Even though biometrics, as such, is a wider term, it is often used for genetical statistics involving studies of plants and animals for assessing their genetic potential in selection trials for genetic material improvement. Such trials are often conducted to develop new better performing varieties or animals with respect to chosen traits. Estimation of genotypic parameters, means, variances, variance components, heritability, genotypic correlations etc. are often computed for such purposes. Performance of varieties and animals under different environments are not the same. Consequently, the stability of performance is often studied along with genotypic-environmental interactions. Population dynamics, the gene frequency estimation in randomly mating populations is another area of study in biometrics. Modelling of biological phenomena, especially, the growth dynamics, insect/pest population dynamics are important areas of study. The true models do not conform to linear forms and one needs to study such relationships in their inherent non-linear forms.

It is needless to emphasize that statistics plays a crucial role in agricultural research whether it is the question of compiling agricultural statistics, or it be conduct of agricultural experiments along with the techniques of drawing valid inferences about treatments/varieties/animals etc. or a variety and animal improvement program.

On the other hand, computers have brought in miraculous changes in the life of modern man. Like the axe was once an extension of human hand, the computer has, now, become an extension of human brain for the modern man. The analogy of a human brain with modern computer is quite interesting, in the sense that today's computer performs quite similar functions as the human brain does. It has almost unlimited memory to store information with possibility of retrieval as per our need, it has logical and arithmetical skills and can take decisions based thereupon, it processes and arranges information like we can do but at a lightening speed. In fact, computer performs better than human beings in some respects as it can work tirelessly, has no memory recall error and can store huge information in its memory in a fraction of a second. The learning curve of computer is unbelievably fast as compared to human beings.

Not so long ago, the computers were termed as white elephants, and rightly so if we consider the huge physical size, the poor processing speed, the exorbitant procurement costs and very expensive maintenance cost of those old generation computers. These were specifically meant to be useful to only the group of scientists engaged mostly in projectile calculations and high speed scientific computations. However, the scenario has changed with computers, now, becoming affordable in price, very powerful capabilities not only in terms of its computing power but the great processing power and the miniaturization in size. What is more important is that, now, the computer technology pervades all walks of human life and has become an essential tool for every one, not limited to the elite only. It is said that in the next generation the computer illiteracy shall be as despicable as the alphabet illiteracy is considered today

In the wake of agricultural development through scientific means, the need for information and data has also increased manifolds. Computers having immense power of information/data processing, storage and retrieval has thus become an important tool for

this purpose. Through computer networking and use of modems etc. it has become extremely useful for communication/dissemination of processed information at lightening speeds. Moreover, the agricultural scientists and others, involved in agricultural research and development, need not depend upon any trained manpower for the processing of their data/information, etc. Developments in computer technology is taking place at mind boggling pace and is bringing about revolutionary changes around. With the advent of very powerful personal computers and user friendly software now, it has become possible for everyone to work on computers sitting right on their desks. Thus the opportunities for computer application in agricultural development have further increased.

The term agriculture in the wider sense includes crop production and its protection, livestock and animal husbandry, dairy, fisheries and related activities such as soil and water management, irrigation and drainage systems, agricultural engineering and post-harvest technology, wasteland and watershed developments, agricultural extension and transfer of technology, credit and cooperation, agricultural marketing, agro-meteorology, environment and forests and many other related areas. In the context of **Panchayati Raj and Decentralized Planning**, information on all these aspects of agriculture is required not only at the national or state level but also at the lower levels such as districts and panchayat levels. Information gathering and storage on multifarious aspects of agriculture in so much detail without the help of computers is not possible.

Agriculture development also requires **formulation, implementation and monitoring** of research programs, backed by extension efforts and also the evaluation of development plans and programs. The databases and information systems relating to these require input from various sources before their processing and dissemination for the planners, policy makers, researchers, administration and other coordinating agencies. Recent advances in computer and communication technology have made computer hardware and software more affordable and have resulted in faster movement of information and its utilization. To expedite communication of data/information between the producers and their users, integrated computer information management systems are being developed using computer and communication technology. Several agencies like NIC, ISRO, ICAR, IASRI and other non-government agencies are engaged in the development of databases and information systems. Updating of databases has also become much easier with the help of computers. Through ARIS (Agricultural Research Information System), ICAR is trying to establish a countrywide network for creating several types of databases such as personnel information system, research projects information system, financial management systems, etc. for the entire Agricultural Research System in the country.

Integrated rural level development and micro-level planning has also been stressed for many years in the successive National Development Plans. Since India has different types of terrain, natural resources, climate, socio-economic levels, administrative set up, cultures, etc., micro-level planning and modeling requires a comprehensive panchayat/village level, spatial and non-spatial databases and integrated information systems. Such databases and integrated information systems can be developed using computers and communication networks.

Computers have played a very significant role in agricultural research. The use of computers for analysis of data pertaining to research experiments is well known. Research planning in agriculture is also a complex process. The research planner must identify a specific problem, define specific objectives, construct hypotheses, mobilize resources for

experiments, disseminate results and continuously re-assess the research results. Computers are extremely useful for these activities.

Statistics plays a major role in planning experiments, conducting experiments and the decision making process in Agricultural Research. Computers have been used for statistical processing of data since long. Analysis of field experiments involves complicated computations which consumes considerable amount of time of the research workers before arriving at a logical conclusion. Now with the availability of Statistical Software packages, the drudgery of computational labor has been almost eliminated and within a very little time, the logical conclusions about the experiments are available. Moreover, multivariate statistical techniques are often complex and so computation intensive that research workers shy away from using these powerful and logically correct techniques on the plea that data analysis would consume lot of time. Statistical software packages have changed the scenario and now agricultural research worker is willing to use any advanced statistical technique for which the package provides for the analysis.

The statisticians have been one of the first few users of computers. In fact, the first commercial computer was used for the tabulation of the US Census data as the time lag between the conduct of census inquiry and the availability of the final results was considerably large. As the computers have advanced over time, these are increasingly being used for the compilation and tabulation of not only the Census data but also of the large scale sample surveys data. The data entry, storage, retrieval in the desired format, making cross classified tables and applying statistical functions and procedures has become all too easy with the availability of powerful computers.

Research management is another area where the use of computers has increased considerably for determining research priorities by allocation of resources for specific commodities or projects. The project management is an important area where computers are increasingly being employed. Optimal allocation of resources is important in multistate, interdisciplinary research programs where activity network can be developed by scientists and administrators on computers to identify the critical research activity, organize these activities in time and formulate the integration of research that can lead to optimal allocation of limited funds/ resources to research programs by assigning benefits, probabilities of success and time and resource requirements to alternative programs. Computers can be used for development of management information systems for supporting such activities.

Computer software for statistical needs has advanced considerably starting with FORTRAN and BASIC programs oriented towards application of a particular statistical tool to the development of comprehensive software packages capable of handling all types of statistical needs, advanced tools and capabilities such as Mstat, Genstat, GLIM, Minitab, SUDAN, SPSS and SAS etc.

Statistics, as a science for studying the techniques for collection, compilation and analysis of data and drawing inferences from the given data set, was supposed to provide tools for extraction of information from the data, the data itself being viewed as numerically expressed facts obtained in a given field of enquiry. If a statistic was sufficient for estimating a given parameter, it was supposed to capture all the relevant information for that parameter. We were, then, in an era of data processing and the computers were classified as electronic data processing (EDP) devices in their right earnest.

The modern situation is much different from that era. We can collect and process the information of every kind, be it in the form of numerical data, categorical data, qualitative attributes or even the textual information, geographic information, maps, colors, patterns, opinions and what not. Information thus transcends the data boundaries and renders the concept of sufficiency, in this sense, less relevant. There is no pressing need for data reduction and data summarization. What is needed is to create a data warehouse (DW) for storing huge mass of data and developing techniques for extraction what we called data mining from such a data warehouse. Statisticians and computer scientists have made much progress in this field and now the software packages even though exorbitantly priced, are available for creation of data warehouse and for carrying out data mining for a given purpose.

**DATA WAREHOUSING:** The emphasis has been gradually shifted from converting data into information to converting information into knowledge and may be, ultimately lead to knowledge converted into wisdom. Data warehousing is an architectural model for the flow of data from operational system into a decision support environment. During 1980's, On-line Analytical Processing (OLAP) and Relational On-line Analytical Processing (ROLAP) tools came into existence for converting data into information. Data warehousing is a repository of subject oriented, integrated, time varying, non-volatile collection of data used primarily in organizational decision making. It is different from a database. By 1990's the data mining and knowledge access tools transformed the information processing scenario by providing influence factors, the data trends and the data pattern recognition for giving an insight into the various associated phenomena.

Creation of a data warehousing involves Extraction, Transformation and Loading (ETL). Extraction refers to the process of obtaining data from various sources, the transformation refers to data integration and cleaning process and finally the loading is actual storage into data warehousing along with its meta data, specifying the concepts, sources, level of aggregation and transformation etc. Meta data is data about data and resides on the top of the data in the data warehouse. Data warehouse also contains a repertoire of various tools for analysis, querying and reporting along with a user system interface. Thus, data warehouse has three layers namely

- Database layer
- Analysis layer
- Presentation layer.

Creation of a data warehouse requires careful planning and designing, effective implementation and regular monitoring of security and performance. The objectives must be clearly spelt out in the planning phase and the requirement analysis for the users must be carried out. The data sources along with quality and periodicity of data availability need to be identified. The network connectivity and data refreshing policies of the organization need to be studied. Accordingly, the data warehouse plan and structure need to be worked out. The design phase involves multi-dimensional data models, data warehouse architecture, its structural schema, physical data model, meta data repository etc. which needs to be decided. For data warehousing normalized schema or star schema is often used. Some of the parameters for the success of data warehousing are, sponsorship of top level executive, understanding user's needs, good quality data, consistent definitions, scalability, use of appropriate tools, continuous end user involvement and regular feedback. Use of different access tools for different users, under-estimation of user demands, inconsistent data definitions, and excessive time spent on data cleaning and data

warehouse building and lack of integration and compatibility are some of the common mistakes in developing a data warehouse.

**DATA MINING:** It is a process of automatically extracting valid, useful, previously unknown and ultimately comprehensible information pattern for taking crucial business decisions. The process of data mining is also known as Knowledge Discovery in Databases (KDD). Data mining tools include association analysis, classification, clustering, modelling etc, using artificial intelligence, fuzzy algorithms, neural networks, multivariate statistical analysis, etc.

Often software tools available for data warehouse and data mining are RDBMS based using Oracle, Sybase and Ingress as backend and using tools for multi-dimensional data base analysis, especially software packages such as Commander, Cognos, Essbase, Red Brick, hybrid packages like Plato and Lightship and statistical analysis packages like SAS.

**INTERNET:** Internet, which started developing during 1990 as a tool for exchanging information has made considerable advancements. There are four components of Internet, namely, Gopher, FTP, Telnet and World Wide Web (WWW). The WWW is the fastest growing service on the Internet and comprises of a Uniform Resource Locator (URL) with hypertext active links in the document including images, text, audio and video. WWW or web publishing is attracting the imaginations of the computer user these days. The static or dynamic web pages with links provide a convenient vehicle for travelling thorough the network. The advertisement on the web is one application. The virtual university possibilities with teaching material on the web and digital library networks provide interesting alternatives to the physical universities and the physical libraries. One can study any course at ones' own leisure and convenience sitting anywhere in the world and at ones' own pace of learning.

Statisticians can hardly ignore the web developments. All the statistical tables, learning materials for various topics are available on the web in spite of the obvious difficulty in expressing mathematical symbols, formula and derivations in HTML language, the language for the web. However, a mathematical markup language is on the anvil and it would not be far away when one could learn mathematical derivation of proofs of well known statistical problems on the web.

Even though technically it is possible to analyze given data sets using statistical packages loading some where else in the cyber space, for business reasons wide availability of such facilities on the web is not there. However, such boundaries are likely to fade out in the future.

**METADATA FOR HETEROGENEOUS WEB RESOURCES:** To overcome the heterogeneity of data sets and information scattered in variety of formats such as text, images, sound, video etc., and the limitations of the search engines for textual searches only, proper organization and cataloguing of the web resources becomes essential. Meta data is a tool for integration of heterogeneous data sets. Meta data supports identification, description and location of network electronic resources so that data catalogue are interoperable i.e., are accessible under one common publicly known interface by hiding their heterogeneity and distribution. As the meta data is data about data sets itself, these represent higher level information describing the content, context, quality, structure and accessibility of a specific data set. It may reside as a header to a resource. HTML

provides for a meta tag for storing meta data information in the web document. Basically meta data describes three components:

- ❖ **Content representation:** basic details about who, where and key linkage
- ❖ **Database description:** data set, model, platform, purpose etc.
- ❖ **Database coverage and availability:** spatial, temporal, limitations, access etc.

Data warehouse makes use of meta data standards for storing and accessing information of various types. Digital libraries also make use of meta data standards for storing information about the documents such as title, author, publication year, pages etc. using Dublin Core Standards.

Besides artificial intelligence, pattern recognition and neural networks also deserve closer examination by the statisticians for applications in their areas of study.

It would thus be obvious that statistics and computers both play a very significant role in agricultural research. Statistics has been using computing devices of all sorts from the beginning, but with the advent of powerful desk computers, the statistical analysis has become very easy. Many times, this ease is interpreted as a substitute for a statistician so much so that the computers are perceived as a threat to the profession of statistics. Of course nothing can be farther from truth than this wrong notion, as no machine can ever replace the expert advice of a good statistician in a given situation. The prudent selection of statistical tools to be applied for a specific problem, weighing the pros and cons of inherent assumptions implicit in each tool, the sequence of application of tools and the drawing of valid inferences, can not be a wise substitute for automated computations and applying tools without considering their limitations.