STATISTICAL SCIENCE IN AGRICULTURAL RESEARCH AND DEVELOPMENT*

P. NARAIN
Indian Agricultural Research Institute
New Delhi.

I feel greatly honoured to have been chosen as Sessional President and invited to deliver the Technical Address during the 47th Annual Conference of the Indian Society of Agricultural Statistics. I express my profound gratitude to the Society for affording me this opportunity. My career interests have been in statistics with applications to agriculture, genetics and breeding and later in research administration. From this background, I propose to share with you some of my thoughts on the nature of statistical activities in agriculture.

Statistical science as we know is concerned with the twin aspects of efficiently collecting observational data with the help of the theory of design of experiments and sample surveys and drawing valid inferences therefrom using the theory of estimation and testing of hypothesis. However, as a scientific activity it is unique in the sense that it does not generate data by itself but needs a field of application to be able to do so. It is therefore at the interface of statistical methods and the field of application that statistical science really exists. Agricultural Statistics is one such interface for which this Society and the Indian Agricultural Statistics Research Institute (IASRI) at New Delhi have made significant and unique contributions since forties and fifties. I have been associated with both of them for a pretty long time. As Director of the latter for a period of over 10 years (1981-1992), I have been actively concerned with the research and training in statistics applied to agriculture.

PAST ACHIEVEMENTS

The fillip to introduce modern statistical methods in agriculture-initiated with the publication of R.A. Fisher's book *Statistical Methods for Research Workers* in 1925 - was given in this country with the setting up of a Statistical

* Technical Address delivered by the Sessional President on 16 December, 1993 at the 47th Annual Conference of ISAS held at S.V. Agricultural College, Tirupati (Andhra Pradesh).

Section in 1930 in the Indian Council of Agricultural Research (ICAR) and the financial support given by the Council to Statistical Laboratory, Calcutta set up around the same time (Panse, 1955). During the next over 60 years, statistical research—both theoretical and applied—were pursued with great vigour over a wide front and earned a place of honour for Indian statisticians in the national as well as international spheres.

By far the most significant contribution, which is now widely adopted not only in India but also in several countries of Asia and Africa, is the method of estimation of yield of crops through crop-cutting experiments in randomly located fields in randomly selected villages of the region under study. A considerable amount of experimental work for judging the optimum size and shape of plots in crop-cutting surveys was done by the Statistical Wing of the ICAR, and the Indian Statistical Institute. The former, on the basis of numerous tests showed that plots of small size give an overestimation of yield. However, recently Verma *et. al.* (1988) reported the results of the study sponsored by Longacre Agricultural Development Centre in London. The study was conducted in different countries of Africa during the course of 1987 to compare the objective method of crop-cutting experiments for estimating crop yields with the subjective method of estimating it by enquiry from farmers soon after the harvest. Their report concluded in favour of subjective method and pointed out that estimates based on square-cuts appeared to lead to overestimation by as much as 30 per cent. This report was presented in a Workshop on Food Supply Information Systems in Africa held at Nairobi (Kenya) in March 1989. At the instance of the Commonwealth Secretariat, London, I prepared an Issue Paper for the Workshop based on the review and commentary on the Longacre report. I reiterated the Indian experience spread over about 35 years which established the superiority of the objective method of estimation of crop yield on the basis of crop-cuts (Narain, 1989). While it is known that small plot size leads to overestimation due to location bias and errors in demarcation of plots, the magnitudes of overestimation obtained in Indian studies (Sukhatme, 1946. Sen, 1967) are found to be much smaller (about 1 to 7 per cent) than the estimate of 30 per cent obtained in the Longacre study. By adopting subjective method based on Farmer's Reports there would be no scope to control non-sampling errors which are apt to be more systematic than random resulting in poor quality data particularly in the long run. In contrast, crop-cutting technique is a time tested procedure having an inbuilt mechanism of controlling the quality of data. The Workshop therefore concluded that Longacre study, being one-time experiment, needs to be repeated before arriving at a final conclusion.

Adhikari (1993) while discussing the social construction of the statistical estimation of crop yield indicates how the controversy due to different approaches to crop-cutting experiments adopted by the ICAR and ISI arose. It is interesting to know that the controversy was due to the differences in the regions in which they experimented. The permanently settled areas of West Bengal where ISI experimented did not have the craft of crop measurement and associated agents, whereas the temporarily settled areas of U.P. and Bihar where ICAR experimented did have the craft of crop measurement and people performing the craft.

Sampling techniques for the estimation of area and production of important plantation crops, fruits and vegetables as also those for the estimation of livestock numbers, products and fisheries were subsequently developed. It may be worth mentioning that although the methodologies of the agricultural surveys differ because of the specific problems of the individual survey,they are similar in the basic approach (Narain and Srivastava, 1993). The emphasis is on the objective approach of measurement on items for which collection of data through other methods like enquiry may lead to sizeable response errors. Further these methodologies take into account the available infrastructure of data collection and analysis which is very essential for the viability and adaptability of the methodology. In fact the methodological research of this kind is a continuous process in which the theoretical and other technological changes must be taken into account simultaneously. A continuous interaction amongst the producers of the data, users and research workers is a basic requirement for this type of research in agricultural surveys.

Apart from practical researches, considerable amount of research in sampling theory has also been carried out by workers at IASRI, ISI and departments of universities including those in the field of agriculture. A unified treatment of sampling theory and its application to large scale surveys was published in the form of a book entitled *Sampling Theory of Surveys with Applications* by P.V. Sukhatme and B.V. Sukhatme in 1954 which has since undergone several editions and modifications. The book, consulted widely by students and researchers, is regarded as a landmark in the progress of Indian contribution in the field of sample surveys.

In the field of design and analysis of experiments, the first report which pointed out the deficiencies of the old type of experiments with inadequate replication and no randomization was that of Vaidyanathan's on *Analysis of Manurial Experiments in India* published in 1934. Consequently, experimental designs such as randomized blocks and latin squares propounded by R.A. Fisher and embodying the principles of replication, randomization and local control and later developments, primarily due to Yates, such as factorial arrangements involving confounding and various incomplete block designs were taken up by the experimental workers. The ICAR encouraged the adoption of modern experimental designs directly by making it a condition for all the research schemes financed by it that the layout plans of the experiments should be approved by the Council's Statistical Adviser. With this emphasis on scientifically planned experiments, there was a rapid spread of the statistical methods in the planning of field experiments and in the analysis of their results. This led to, at the instance of the Council, the publication of a book *Statistical Methods for Agricultural Workers* by V.G. Panse and P.V. Sukhatme in 1954 which has since been revised. It may however be pointed out that the role of agricultural statisticians in the advisory activities has now been relegated to the background as nowadays agricultural scientists claim to have already a first hand knowledge of statistics to handle designing and analysing their own experiments.

Compared to crops, the introduction of modern experimental design in animal husbandry has been rather a slower process due to inherent limitations of experimental material such as the cost of maintaining animals and long duration of experiments as in cattle and buffaloes. Apart from the simpler designs like randomized block and latin squares, more involved ones such as switch back and switch-over designs are often employed. A useful publication in this connection brought out by this Society in 1975 is that of *Statistical Methods in Animal Sciences* by V.N. Amble.

A critical statistical analysis of ten year's data collected in a goat breeding project at Etah in U.P. by Sukhatme in 1944 became the landmark in focussing attention of animal breeders on the need for a continued statistical appraisal of the results of breeding. In the Etah project, it was shown that year to year improvement recorded in the herd was not due to genetic improvement of the stock through selection but due to non-genetic factors. Later on, the results of the analysis of cattle breeding data on Red Sindhi herds at Bangalore and Hosur conducted by Amble, Krishnan and Srivastava in 1958 paved the way for development of progeny testing programme for the improvement of village cattle through the use of sires tested at the farms. Several plant breeding studies such as the estimation of genetic variability in self-fertilized crops like cotton (Panse, 1940a), development of genetic models and effective number of factors for selection of yield in cotton (Panse, 1940b) were also conducted. These studies in plant and animal breeding laid the foundation of a new field of activity of *Statistical Genetics* which is one of its own kind and conducted at few institutions like IASRI. It is essentially interdisciplinary in nature requiring an integrated approach placing due emphasis on both the principles of statistics as well as genetics. A unified treatment of its theory from this integrated point of view has been described recently in a book entitled *Statistical Genetics* by P. Narain and published by Wiley Eastern Ltd., N. Delhi in 1990 which has been reprinted this year.

The above description on the achievements in the field of statistics as applied to agriculture raises an important question as to how agricultural research and development have gained from them. The most direct gain seems to have come from the sample survey work for the assessment of reliable food production which is needed for planning purposes. Besides, several sample surveys conducted on the farmers' fields have helped in monitoring what farmers actually do, which give a feed-back to the agricultural administration and extension agencies for eliminating the poor agronomic practices etc. In the field of experimental research, several newly evolved designs have helped in reducing the cost of experimentation. It is claimed that if certain incomplete block designs are adopted in the place of complete randomized blocks in crop experiments, there can be as much as 70 per cent saving in the use of land and labour for varietal trials (Gower, 1988). Apart from such direct impact of statistical science, more often than not, statistics has played its part indirectly by encouraging the efficient use of limited research resources and discouraging the unjustified research.

FUTURE TRENDS

Due to the vast increase in computer power that has become available in the recent decade, the whole scenario of statistical data analysis seems to have changed. Besides, new exciting developments in agriculture have taken place which the statisticians have to follow for solving the problems of the real world such as for instance biotechnology. Time does not permit me to indicate the possible statistical developments in all the emerging new areas in the field of agriculture and allied disciplines. I shall therefore restrict myself to a few of those which need urgent attention of the agricultural statisticians. These are:

1. *Statistical prediction*
2. *Statistics in biotechnology*
3. *Statistics, information technology and agrarian society*
4. *Computer-intensive statistical methods*

Statistical Prediction

Statistics is concerned with the development of scientific methods of induction from unpredictable quantitative data. The statistical inference attempts to minimize the arbitrariness of induction by evoking deductive methods to a large measure. This is the background to the development of the classical statistical science in terms of the theories of estimation, testing and decision making by Fisher, Neyman and Pearson and Wald. In particular, Fisher (1925a), while talking about the estimation stated the objective of the statistical method as reduction of data in such a manner that the whole of the relevant information contained in the data is retrieved while excluding all the irrelevant information. This led him to the issue of specification which he dealt with by specifying a distribution of the observed characteristics. For instance, if we have a sample of yields corresponding to different doses of fertilizer in an agricultural experiment, we first specify the relationship between the expected yield Y and the fertilizer x by

$$Y = a\,[1 - \exp\{-k\,(x - b)\}],$$

Then we take the observed yields y as normally distributed about Y as mean and variance σ^2. We then have the problem of estimating a, k, b and σ^2 in the best possible manner. In terms of a simple model, we state $y = Y + e$, where the errors e become normally distributed with zero mean and variance σ^2. But we never think about estimating e itself, and go to the estimation of the second degree statistics σ^2. If we estimate e, we can predict Y for future experiment. It seems therefore that a more general approach to the whole issue is to talk about *prediction* in the statistical sense. In fact the purpose of any agricultural experiment is *prediction* i.e. say out of the two varieties being compared which one shall we use in the future. In recent times therefore efforts have been made to develop procedures of estimating *random effects*. One such procedure is *Best Linear Unbiased Prediction* (BLUP) which has been extensively used in the field of animal breeding. There are similiarities between BLUP and other techniques like *Kriging* and *Kalman filter* used in geostatistics and control theory respectively. Interestingly, although BLUP was developed via a frequentist approach to statistics, it has a Bayesian interpretation too.

It seems the whole issue can be given a unified treatment by developing a general prediction theory (Harville, 1990). We consider predicting the value of an *unobservable* random variable x based on the value of a $n \times 1$ *observable* random vector y where the joint distribution of x and y has first and seond moments denoted by $\mu_x = E(x)$, $\mu_y = E(y)$, $\sigma_x^2 = Var(x)$, $\sigma_{yx} = Cov(y, x)$ and $V_y = var(y)$. We assume that μ_y belongs to a known vector space and that μ_x is a known linear combination of the elements of μ_y. This means $\mu_y = \mathbf{X}\,\beta$ and $\mu_x = \lambda^T\beta$ where β is a $p \times 1$ vector of unknown parameters, $\mathbf{X}$ is a $n \times p$ known matrix of rank p^* and λ is a $p{\times}1$ known vector that is expressible as $\lambda = \mathbf{X}^T\mathbf{k}$ for some vector $\mathbf{k}$. The quantity σ_x^2 and the elements of σ_{yx} and V_y are assumed to be known functions of an unknown parameter vector θ whose value is restricted to a known set Ω and V_y is assumed to be non-singular (for all $\theta \in \Omega$). Apart from the linearity of the mean structure, these assumptions mean that σ_x^2, σ_{yx} and V_y are unrelated to μ_x and μ_y. When we take the special case with $\sigma_x^2 = 0$, x equals $\lambda^T\beta$ with probability one. The problem of predicting the value of x is then equivalent to that of inference about fixed effects $x = \lambda^T\beta$ in the classical sense.

To give one example of the problem of prediction, the evaluation of the breeding value of an individual can be regarded as a statistical problem of prediction- to predict an unobservable random variable (the breeding value) with the help of a set of observed random variables (the averages of the phenotypic values of the concerned relatives). We may be interested in predicting the breeding (genetic) value of a bull with the help of observable records of a given number of progeny of the bull. The form of the joint distribution of records of progeny and the genetic value of the bull is not known as well as the first moment of the distribution is also not known. But only the second central moment is known. In such a case BLUP, given for the first time by Henderson (1963), can be used, In this method, the linear function of the records which has the same expectation as the genetic value to be predicted and which, in the class of such functions, minimizes the average of the squared errors in prediction, is the desired BLUP. When the joint distribution is taken as bivariate normal with numerically known values of the first and second moments, the conditional mean of the genetic value, given the records of progeny, gives the accuracy of the progeny test which is found to depend on the number of progeny and the heritability of the trait. Introduction of auxiliary traits in such problems improves the accuracy of the progeny test (Narain, 1985; Narain and Kaur, 1993).

Statistics in Biotechnology

The discovery of the fascinating biological science of genetics is regarded as by far the greatest scientific accomplishment of the twentieth century. With the advent of molecular biology and genetic engineering, noval life forms not found in nature could be created. Through recombinant DNA technology, individual genes can be isolated, cloned and transferred across species boundaries from plants, animals, microorganisms or viruses. Not only this, using protoplast fusion technique, it has been possible to transfer large segments of genome from one bacterium to another and to create hybrid organisms that can be propogated as for instance tomato-potato hybrid. The introduction of these new technologies has made significant impact on some of the statistico-genetic principles used in breeding for improvement of domesticated plants and animals. We discuss below two of them:

Restriction Fragment Length Polymorphism (RFLP)

Most of the traits of economic significance in plants such as yield exhibits quantitative differences whose mode of inheritance is complex in nature. Usually a large number of genes scattered over different chromosomes with each of the gene contributing a small positive or negative effect to the trait and having a large environmental component at the phenotypic level are involved. So far we had no means to identify the effect of individual genes and follow them individually through generations as we do in the case of Mendelian genes. The emergence of the technique of *Restriction Fragment Length Polymorphism* (RFLP) has now opened up new possibilities for the identification of quantitative trait loci (QTL), by means of correlation between the trait and the specific RFLP markers. RFLP maps can be used to select desired gene combinations in an indirect manner from a breeding population, usually called Marker Assisted Selection (MAS). It has also been helpful in selection of desirable genotypes in early generations from a segregating population and is being used extensively in breeding improved varieties of a number of crop plants across the world. The first most notable achievement in this direction has been in tomato (Patterson *et al* 1988) where QTLs governing fruit mass and fruit pH have been identified and selected for high solid content. In India, RFLP studies have been initiated and are being used for molecular- mapping and character-tagging studies in mustard (*Brassica spp*), chick-pea (*Cicer arietimum*) and rice at the Biotechnology Centre, I.A.R.I. New Delhi, National Chemical Laboratory (NCL), Pune and the International Centre for Genetic Engineering and Biotechnology (ICGEB), New Delhi.

The experimental method for analysing QTLs in plant species involves crossing two independently inbred parental strains that are phenotypically widely different from one another in respect of several quantitative traits, and then analysing the genome of a number say (50 to 100) of the recombinant hybrids using a complete series of mapped and ordered RFLP probes to determine in each of the hybrids which segments have been inherited from one parent and which from the other. The phenotype of each hybrid with respect to the trait is then compared with the RFLP-based genome-analysis data to determine statistically which segments of the genome contribute significantly to the ultimate phenotype and are therefore likely to harbour QTLs.

The statistical approach for the detection of QTL near a genetic marker locus involves comparing the phenotypic means for two classes of progeny: those with marker genotype *AB* and those with marker genotype *AA*. The difference between the two means gives an estimate of the phenotypic effect of substituting a *B* allele for an *A* allele at the QTL. For testing whether the inferred phenotypic effect is significantly different from zero, one uses the linear regression of phenotype (P) on genotype (G) given, for the ith individual by

$$P_i = a + b\, G_i + e_i$$

where G_i is encoded as a (0, 1) — indicator variable equal to the number of *B* alleles, e_i is N (0, σ^2), for i = 1, 2, . . ., n and a, b are unknown parameters, b being the phenotypic effect of single gene substitution at a putative QTL. It is assumed here that the QTL is located at the marker locus itself so that G_i is known with certainity. The linear regression solutions (a, b, σ^2) are in fact maximum likelihood estimates (MLEs) L (a, b, σ^2) based on normal distribution. The constrained MLEs are obtained on the assumption that b = 0 i.e. on the hypothesis that no QTL is linked. In the case of back-cross with *A* as recurrent parent, this MLE is $L(\mu_A, 0, \sigma^2_{B_1})$ where μ_A is mean of *A* parent and $\sigma^2_{B_1}$ is phenotypic variance in the back-cross population. The evidence for a QTL is then summarized by the LOD score given by

$$\text{LOD} = \text{Log}_{10}\left[\frac{L\,(a, b, \sigma^2)}{L\,(\mu_A, 0, \sigma^2_{B_1})}\right]$$

When LOD score exceeds a pre-determined threshold T, a QTL is declared to be present. For large samples, 2(LOD) is asymptotically distributed as $(\log_{10}e)\,\chi^2_1$ where χ^2_1 is chi-square with 1 d.f.

With RFLPs scored throughout the genome we consider flanking marker genetic models in which a QTL lies between linked codominant marker loci. The maximum likelihood solution in such a case is in fact a linear regression problem with missing data i.e. none of the independent variables (genotypes) are known; only their probability distributions are available. Hence standard computer programmes for linear regression cannot be used. For this purpose, Lander and Botstein (1989) adopted the *EM* algorithm.

Multiple Ovulation and Embryo Transfer Technology (MOET)

Breeders of dairy cattle and buffaloes have exploited Artificial Insemination (AI) for commercial and genetic advantages. The advent of biotechnology has led to the emergence of new techniques like embryo sexing, embryo cloning and embryo transfer which allow production of bulls out of a few *top cows* which can result in genetic improvement in milk yield. Under the *Multiple Ovulation Embryo Transfer Technology* (MOET), donor high producing cows and local low genetic quality recipient cows are synchronised by administering with suitable hormones and at the appropriate stage, the donors are superovulated. The ova are fertilised through AI from a valuable sire. The embryos are finally recovered non-surgically and transferred to the synchronised recipient cows which act as surrogate dams. This technique provides introduction of genes of high quality genetic stocks both from sire as well as from dam. In addition, since the foetus receives blood supply from the local surrogate mother, natural immunity against some local diseases gets transmitted to the foetus.

In a traditional progeny-testing programme with AI we normally test 150 young male cattle bulls in testing herds per year. For an efficient plan, we test each bull on about 50 daughters and use the best two bulls for one year to breed bulls and the best ten bulls for three years to breed cows. It can be shown that such a programme leads to an estimated genetic change of the order of 0.1 phenotypic standard deviation units per year. However in practice the rates of realised genetic change in milk yield are much lower being of the order of 0.005 to 0.0075 of the mean. If we involve embryo transfer either from one year old females (juvenile scheme with generation interval of 1.8 years) or from females after one lactation (adult scheme with generation interval of 3.7 years) with the use of males at similar ages, increased rates of genetic change can be achieved (Nicholas and Smith, 1983). Though selection is less accurate than in the conventional progeny testing, the annual rate of genetic improvement can be increased and even doubled. If the number of transfers is restricted and the inbreeding rate is limiting, the adult scheme for both sexes is preferred. Because of the relatively small number of animals involved, greater control over recording, breeding and selection should be possible, leading to a larger proportion of the possible genetic gains being realised in practice. The problem of optimisation of the breeding programme using embryo transfer was discussed by Robertson (1983) in a seminar delivered at the IASRI, New Delhi.

In India, success has been achieved in producing MOET baby- cows particularly due to the joint efforts made by the I.A.R.I., Pusa and the National Institute of Immunology, New Delhi under a joint research programme in which Haryana cows are used as recipients and pure Holstein Friesians and cross-bred cows are used as donors. However, the statistical aspects of the technique are yet to make their impact.

Statistics, Information Technology and Agrarian Society

Modern computer power has completely changed our outlook on the use of statistics for decision making. The data generated at the ultimate level like a district in India in various sectors such as agriculture, power, health, family welfare, industries, education, employment, labour, transport, communication, forest, roads etc. until recently used to be handled manually and consolidated progressively at the districts, States and Centre. Such a manual process not only lacked comprehensiveness and accuracy but also involved inevitable delays and hampered decision making capability in day to day administration as well as in the planning process. This is now fast changing with the advent of information technology which involves application of computer as well as communication technologies. The ability of a computer system to store and process considerable volume of data, to carry out complex analysis at high speed, to integrate effectively the man-machine interface for decision making functions and to adapt, through software, to changing system requirements, means that nowadays it is possible for the same computer to address all aspects of data gathering, information processing, process control, on-line optimisation, operations control and even real time scheduling and production planning functions. Coupled with latest communication technology, we have entered into an era of information-oriented network for handling and transporting a variety of information in diverse ways.

In important sectors like agriculture, animal husbandry, fisheries and forestry, sound statistical methods involving random sampling techniques, as already discussed, are adopted to provide reliable estimates. However, until recently most of the work has been undertaken manually with long delays, outmoded statistics for planning and in several cases lacking in comprehensiveness. Only recently, National Informatics Centre (NIC) of the Government of India has established *NICNET*-a computer-communication network, which serves as a bridge between the State Governments and the Central Government on the one hand and the State Governments and their district administration on the other (Seshagiri, 1991). Under this scheme, the basic data are captured by a systems approach at the district level. NIC has launched *District Information System Programme (DISNIC)* for this purpose. The NIC with the help of State Governments and through pilot projects, has developed DISNIC database templates for as many as 26 sectors including agriculture.

The information system *DISNIC-Agriculture* has been developed for district-wise collection and dissemination of agricultural data. This is a user friendly, menu driven package developed using Foxbase + in XENIX environment for implementation on the computers in all the districts and state capitals. Information on agricultural population, wages, rainfall, area under crops, fertilizer and pesticide, irrigation, crop-wise production, prices, etc. are included in the package. By pressing a few keys, the user can obtain any of the information for feeding or reading the data into it. The district-wise data bases can thus be created easily by any non-computer professional.

KRISHINET

Although several information systems have come up in different sectors, none so far has come up in the field of agricultural research and education. Narain (1990a; 1992; 1993) and Narain and Gopalan (1992) have put forth a blueprint of *KRISHINET*-a nation-wide *Agricultural Research and Education Information System.*

At present the task of agricultural research and education is undertaken in various institutes and centres of ICAR and State Agricultural Universities (SAUs). While around 6,000 to 7,000 agricultural scientists are employed in the ICAR system, about 25,000 agricultural scientists are engaged in research, education and extension in different SAUs. Due to the efforts of these scientists, research data information have accumulated in the past several years and more and more are produced each year. Because of widely dispersed locations of the institutes/universities in a vast country like India, rapid scientific communication of such data/information amongst the scientists is very meager. It is for this reason that a comprehensive information system in the form of a computer network service like *KRISHINET* is very much required.

Mechanism will need to be created by which the research centres in the institutes/universities will be provided with comprehensive current data/information in all fields of agricultural research. The research information will flow from individual research scientists of the individual research centres of their institutes/universities which will need to be exchanged between all concerned. Such data will have to be input in the local computer systems for necessary processing, merging and output information. Once the comprehensive information service becomes operational, current data from all the centres will flow into the system. Standardised formats will have to be evolved for input data generated at a large number of diverse centres for networking of such information. SAUs will have to be interconnected as well as connected with the district and state systems of the NIC. In other words, the *KRISHINET* will have to be integrated with the existing communication net-work *NICNET* of the NIC or any other central network in the country like *INDONET* of CMC. Wherever such communication does not exist, ICAR will have to provide for the same. In a similar manner, the various Institutes, national research centres and project directorates of the ICAR will be net-worked.

Through the gateways to be provided by central net-works (*NICNET or INDONET*) these two big agricultural systems viz. ICAR and SAUs will have to be inter-linked to another ICAR institution like IASRI, New Delhi which is centrally located as well as have the necessary expertise in computer research and education. The IASRI system will act as the main host and will have the master earth station for accommodating the micro-earth stations installed in different ICAR Institutes and SAUs. It will also have to be linked with the central net-work Headquarters at New Delhi. In fact the type of communication-based computer system that can be considered for *KRISHINET* is the resource sharing net-work interconnecting processing sites operated and used by numerous organizational entities. The concept is that of a super net-work in which any individual user is permitted to share a part or all of the computational resource at any of the interconnected processing sites.

Computers in Agrarian Society

One of the alarming situation in the urban-rural fabric of our country seems to be the traffic from rural to urban areas. More and more people particularly younger people are shifting from rural areas to urban areas in search of better employment, better way of life and better comfort. This has led to imbalance in the growth of agriculture in our system. If no proper remedy is made, it may lead to serious consequences, because what we eat every day is produced in the innumerable fields in the countryside. As such, efforts are needed to take science and technology to the farming community themselves. Several transfer of agricultural technology programmes are therefore being undertaken to demonstrate to the farmers the benefits of science and technology in their own fields. In a similar manner, the benefits of a computer need to be brought home to the agrarian society in the rural sector (Narain, 1985a). Computers can be very helpful to farming community in providing them very valuable and up-to-date information on agronomic and related aspects in relation to their crop production. This; of course, would require considerable efforts to generate data on agriculture, up-date them periodically, analyse them and disseminate the resulting information in a useful format to the farmers for decision making in their own occupation. This calls for taking the information technology to the agrarian people.

In some Western countries, such as for instance in Canada a Telidon system is already in vogue which helps the farmers immensely. There, the weather is monitored in relation to crop production all over the country. The Department of Land Resource Research Institute in the Research Branch of Agriculture, Canada makes available information on daily weather bulletin, soil moisture and its evaluation through real-time via a VAX computer. This has led to the *Grassroots* service which is a comprehensive agricultural information service that gives the farmer instant access to the kinds of information necessary for the profitable and successful operation on the farm. The Telidon system is a two-way television. A farmer can have, instantly, at his fingertips, accurate local weather, commodities and livestock markets, grain newsletters and analysis, government bulletins, new product developments and much more. All are available at *Grassroots* always up-to-date. The system is such that using a simple keypad, any of the thousands of pages of information stored in the Central Computer is made available on the television screen. The kind of information and the amount that can be stored in the Central Computer is practically unlimited. There are thousands of pages with more being added every day. Any information that helps the farmer in planning or managing his operation on the farm can be made available on this system.

Although such a system has not reached India, Teletext and Videotext advances in computer and communication technologies are being introduced in the more elite urban areas and for such uses as airlines reservation, general public issues, entertainment in big cities and the like. Just as television, when introduced in fifties was only limited to big cities, has now gone over to a much larger area in the country, it is hoped that the computer and communication technologies will catch up soon for agrarian society because of its immense potential for the use and need of the farming community.

Computer-intensive Statistical Methods

The revolution in the field of electronic computation has led to the development of new statistical theory and methods in the 1980s, particularly the upsurge of computer-intensive statistical methodology such as *bootstrap methods, non-parametric regression, generalized additive models* etc. The most significant point about these methods is that fewer distributional assumptions are required for their application than the corresponding classical statistical methods. The replicability of the results of the experimental scientists can therefore be better ensured. For instance, in most of the cases of the classical methods, normal distribution is invariably assumed whereas in practice it need not necessarily be true so that conclusions obtained from such methods are often approximate or even get vitiated unless these methods are robust. No wonder then that on repeating the experiment, we get different results. At a more practical level, computer-intensive techniques have been helpful in developing sampling strategy for spatial data, particularly in the context of modem Geographical Information Systems (GIS). We discuss below some of these computer-intensive techniques.

Bootstrap methods discussed by Efron and Tibshirani (1991) are basically simulation methods conducted on high-speed computers and aimed at generating new data sets from the observed original data set. The term *bootstrap* -derived from the old saying about pulling yourself up by your own bootstraps-reflects the fact that the one available sample gives rise to a large number of other samples. It is essentially a resampling technique on computer to extract as much information as possible from the data on hand. Suppose a researcher has only a single data set consisting of a given number(n) values. By the bootstrap method the computer copies each of the values say, a billion times, mixes them thoroughly, then selects samples consisting of n values each. These samples can then be used to estimate any statistic and its standard error - in fact one can get the whole sampling distribution of the statistic. For instance, when the data set happens to come from a long-tailed probability distribution, a *trimmed mean*, where a certain percentage of observations are deleted from the lower and upper ends of the ordered data set, provides a much more accurate statistic than the usual mean. Since we do not know *a priori* whether the true distribution is long-tailed or not, it is preferable to use trimmed means. Suppose we have the following data set of 9 observations as an ordered random sample from say 164 observations

-21.0, 3.25, 10.75, 13.75, 32.50, 39.50, 41.75, 56.75, 80.0

The ordinary mean is 28.58 with an estimated standard error of 10.13. The estimated confidence interval is then 28.58 ± 10.13z where z = 1.645, 1.960, 2.576 relating to the areas under the normal curve for the respective 90,95 and 99 per cent confidence intervals. When we take 25 per cent trimmed mean, it is obtained as 27.81. For obtaining the standard error of this estimate, we do not have any algebraic formula like in the case of ordinary mean. A number of bootstrap samples (B) of size 9 are therefore drawn independently with replacement from the original data set using random number generator on computer, the 25 per cent trimmed mean is calculated for each sample and an *empirical* standard deviation of these sample means is worked out to obtain the bootstrap estimate of the standard error for 25 per cent trimmed mean. For B = 25, this estimate is 12.44 whereas for B = 200, this is 10.70. The randomness in the bootstrap standard errors comes from the use of a finite number of samples. It has been found that this is negligible when the number of samples is greater than 200. The setting up of confidence intervals with bootstrap is, however, not that simple.

The above simple algorithm of bootstrap sampling can be applied in several other cases like regression, time series, stratified sampling, multivariate structure etc. The algorithm is perfectly general. For instance, in the case of non-parametric regression between two variables y and x, *loess* — a computer-based fitting method—instead of usual least squares is used. In this method, no model is assumed over the entire range of the independent variable x. Instead, a series of *local* regression curves for different values of x (the target point) is fitted, in each case using only data points near the target x - value of interest (say a moving window of 20 per cent of the points). The process is repeated for all possible target points to obtain a non-parametric regression. The bootstrap algorithm is applied as before to obtain the bootstrap *loess* curves giving the standard errors of the predicted values at each of the target points.

GIS-assisted computer-intensive sampling strategy

Geographical Information Systems (GIS) are powerful computer-oriented software tools that bring together the task of storing and retrieving large quantity of geographical data, the task of analysing them and the task of drawing any kind of map about them. The GIS can handle various types of spatial information through their geographic coordinates and therefore can be very useful for design and processing of surveys based on area sampling. It is used for agricultural and land use planning, forestry and wild life management, geology and archaeology. The typologies of spatial data can be points, lines, areas or polygons and surfaces. Data arranged in areas are encountered in physical geography when we deal with remotely sensed data like satellite imagery and maps. Here a map is made up of a set of measurements on the character arranged on the regular array of quadrants i.e. the pixels of the resulting image. The spatial data, thus collected, can be stored in a GIS either as *raster* (the image is represented on a regular square lattice grid) or as *vector* in terms of the cartesian coordinates of the origin and destination points in a continuous spatial domain. These two formats are however interchangeable but with loss of information if we go from vector to raster format.

Area sampling is often used in agricultural surveys, particularly when direct sampling of individual units of the population is too expensive or even not feasible or else the complete frame is not available. Various probabilistic strategies are used for this purpose. For instance, text-books on sampling describe techniques as simple random, systematic, systematic unaligned and two-stage or stratified. The selection of areal units is made as if the generating process which underlies the observed values is constituted by a sequence of random variables which are independently and identically distributed (i.i.d.). The probability of selecting one unit is assumed to be independent of that of having selected another one. This however does not take into account the spatial correlation which invariably exists in spatial data and need therefore be exploited in the sampling design. Moreover, the existing sampling techniques treat irregularly shaped data as if they are regular grids of quadrants which could be misleading in many respects. Recently, a DUST (dependent areal units sequential technique) technique of sampling has been developed (Arbia 1993), wherein the sampling procedure is "draw-by-draw" technique with variable inclusion probability at each step. We have a target variable X observable in N non-overlapping areal units within a study area which is related to the auxiliary variable Y. For instance, Y could be a satellite image and X could be ground truth for agricultural and land resources surveys. There are then three steps in the DUST technique.

Firstly, the spatial correlation (β) in the proxy variable Y is estimated at various spatial lags ($\beta_1, \beta_2, \ldots, \beta_n$). The procedure could be simplified by estimating β_1 and then fitting a plausible distance decay models for higher order correlation like $\beta_k = \beta_1^k$. Though the procedure could be computationally very heavy even for small N, it could be easily accomplished in a GIS which has built-in routines for neighbouring search of common boundaries between polygons. In the second step, stationarity of the various order correlations is tested usually by the moving sampling techniques. In the last step, the spatial correlation of the proxy variable Y is employed to assign drawing weights to the individual sample units in a prescribed way. If $\beta = 0$, we locate samples randomly as in the simple random case. If not, the sample is drawn sequentially by assigning a weight varying at each step. The first unit gets a weight of unity, the second $(1 - \beta_1^{d_{12}})$, the third $(1 - \beta_1^{d_{13}})$ $(1 - \beta_1^{d_{23}})$ and so on, the nth unit getting the weight $\prod_{i=1}^{n-1} (1 - \beta_1^{d_{in}})$. In this scheme, n is the sample size and d_{ij} is the distance between the ith and jth area unit measured in terms of physical distance between centroids or in terms of the order of neighbourhood. Such a sampling scheme results in increasing probability of an unit being drawn with increasing distance, from the areas already sampled. After a certain distance, β vanishes and (1 – β) equals one for each unit implying random choice of units. The DUST sampling strategy shows that the use of GIS for computer-intensive approach to sampling spatial data can result in a better allocation of resources leading to higher levels of accuracy of the estimates at a given sample size or alternatively to smaller sample sizes and hence reduced cost at a constant level of accuracy.

CONCLUSIONS

Finally, I hope the role of statistical science in agricultural research and development which I have briefly presented, would indicate immense opportunities for an agricultural statistican engaged in theoretical or applied research and eager to serve the needs of farming community. It is apparent that potential developments in agricultural statistics research during nineties against the background of changing emphasis in food production due to advances in the frontier areas of biotechnology, information and space technologies etc. are going to be much more demanding than hitherto. There is, therefore, a vital need, on the part of statisticians, to specialize in particular areas of application and become fully involved in them. This Society can be of great service in such an endeavour.

REFERENCES

Adhikari, B.P. (1993). Social construction of the statistical estimation of crop yield. (Unpublished).

Amble, V.N., K.S. Krishnan and J. Srivastava. (1958). Statistical studies of breeding data of Indian herds of cattle. Indian J. Vety. Sci. **28**: 33.

Amble, V.N. (1975). *Statistical Methods in Animal Sciences*. I.S.A.S., N. Delhi.

Arbia, G. (1993). The use of GIS in spatial statistical surveys. International Statistical Review. **61**(2):339-359.

Efron, B and R. Tibshirani (1991). Statistical data analysis in the computer age. Science. **253**:390-395.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd., Edinburgh.

Fisher, R.A. (1925a). Theory of Statistical Estimation. Proc. Camb. Phil. Soc. **22**: 700-725.

Gower, J.C. (1988). Statistics and Agriculture. J. Royal Statist. Soc. **A151**:179-200.

Harville, D.A. (1990). BLUP (Best Linear Unbiased Prediction) and Beyond. In: *Advances in Statistical Methods for Genetic Improvement of Livestock* (Eds. D. Gianola and K. Hammond) : 239-276. Springer, New York.

Henderson, C.R. (1963). Selection index and expected genetic advance. In: *Statistical Genetics and Plant Breeding*. (Ed. W.D. Hanson and H.F. Robinson) Nat. Acad. Sci., Nat. Res. Council, Publication 982, Washington, D.C.

Lander, E.S. and D. Botstein. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. **121**:185-199.

Narain, P. (1985). Progeny testing with auxiliary traits. Biometrics, **41**:895-907.

Narain, P. (1985a). Computers and the Agrarian Society. Computer Age.3(6):43-45.

Narain, P. (1989). Methodological issues in estimation of crop prodution in African countries. In: *Proc. Workshop on Food Supply Information Systems*, Nairobi (Kenya), 13 to 17 March. Commonwealth Secretariat, London, UNSO, N. York and EUROSTAT, Luxembourg: 1-40.

Narain, P. (1990). *Statistical Genetics*. Wiley Eastern Ltd., N. Delhi. Reprinted in 1993.

Narain, P. (1990a). Information Input in Agriculture. In: *Technology Blending and Agrarian Prosperity*. Eds. J.P. Verma and Anupam Varma., IARI, N.Delhi:61-70.

Narain, P. (1992). Information Technology and Official Statistics in India. *Proc. Third Independent Conference of International Association for Official Statistics,* Ankara (Turky), 22 to 25 September. ISI: 1-5.

Narain, P. (1993). Information Technology and Official Statistics in India. J. Official Statistics - An International Review published by Statistics Sweden. (In press).

Narain, P. and R. Gopalan. (1992). Information Technology. In: *Proc. Nehru Centenary Symposium* held at New Delhi, April, 1989, FISAST:1-18.

Narain, P. and A.K. Srivastava. (1993). Methodology of Agricultural Surveys in India. In: Proc. *International Conference on Establishment Surveys* (ICES) Buffalo, New York, June 27-30. American Statistical Association, :1-5.

Narain, P. and A.P. Kaur. (1993). The accuracy of progeny testing with binomial auxiliary traits. Animal Production, Journal of the British Society of Animal Production. (In press).

Nicholas, F.W. and C. Smith. (1983). Increased rates of genetic change in dairy cattle by embryo transfer and splitting. Anim. Prod. **36**:341-353.

Panse, V.G. (1940a). The inheritance of quantitative characters and plant breeding. J. Genet., **40**:283.

Panse, V.G. (1940b). A statistical study of quantitative inheritance. Ann. Eugenics, **10**:76.

Panse, V.G. (1955). *Thirty years of Statistics in Agriculture in India.* ICAR Review Series No.**27**:1-28.

Panse, V.G. and P.V. Sukhatme. (1954). *Statistical Methods for Agricultural Workers.* ICAR, N. Delhi.

Patterson, A.H., E.H. Lander, J.D. Hewitt, S. Peterson, S.E. Lincoln, S.D. Tanksley. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphism. Nature. **335**:721-726.

Robertson, A. (1983). Optimisation of breeding programmes for milk production using embryo transfer. Seminar delivered at IASRI, N. Delhi on December 14.

Sen, S.R. (1967). *Report of the Technical Committee on Crop Estimates.* Planning Commission, Govt. of India: 1-127.

Seshagiri, R. (1991). Informatics as a tool for development: A long term perspective. In: *Glimpses of Science in India*-Ed. U.S. Srivastava, Diamond Jubilee Vol., The National Academy of Sciences, India: 384-408.

Sukhatme, P.V. (1944). Statistical study of a breeding experiment with goats. Indian J. Vet. Sci., **14**:167.

Sukhatme, P.V. (1946). Bias in the use of small size plots in sample surveys for yield. Nature, **157**:63.

Sukhatme, P.V. and B.V. Sukhatme. (1954). *Sampling Theory of Surveys with Applications.* I.S.A.S., N. Delhi and Iowa State University Press, Ames. Iowa, USA.

Vaidyanathan, M. (1934). *Analysis of manurial experiments in India.* ICAR, N. Delhi.

Verma, V., T. Marchant, C. Scott (1988). *Evaluation of Crop-cut Methods and Farmer Reports for estimating Crop production: Results of a Methodological Study in five African Countries.* Report issued by Longacre Agricultural Development Centre Limited, London.