Agricultural Statistics in the Twenty-first Century[1]

Jagdish S. Rustagi
Department of Statistics, The Ohio State University, Columbus, Ohio, USA

SUMMARY

Statistical sciences have been applied to the scientific method contributing to the advances of many natural, biological, physical and social sciences. Improvement in manufacturing in industry and growth in agriculture has been accomplished through statistical techniques in modern times. Important advances in quality improvement have been achieved in computer and high technology industry through statistical techniques developed in the twentieth century. It is expected that new techniques such as computer intensive methods will be increasingly used in agricultural research and practice. In addition, many new techniques developed in this century such as those of jackknife, bootstrap, data mining, neural networks, meta analysis, spatial data analysis, Baysian methods, Gibbs sampling and Markov Chain, Monte Carlo etc., will be of great potential applications in agriculture. Just as Ronald A. Fisher, P.V. Sukhatme, V.G. Panse, G. Mendel, etc., made pioneering developments in the twentieth century in agricultural statistics, the twenty-first century will find its own heroes who will further advance the frontiers of agricultural statistics.

1. Introduction

The twentieth century has seen the flowering of the statistical sciences and their applications to other natural, physical, social and environmental sciences. Statistics has penetrated into business, government and industry like no other field. The decision-makers of today depend heavily on the use of statistical techniques. Sir Ronald A. Fisher is credited with a major role he played in this development. Essentially it is he who put statistical sciences on the firm foundations in the early part of the twentieth century. Many others, such as A.N. Kolmogorov, Karl Pearson, E.S. Pearson, Jerry Neyman, Abraham Wald, Harold Hotelling, C.R. Rao and P.V. Sukhatme made prominent advances in the statistical sciences.

---

1 Invited talk given at the 51st annual meeting of the Indian Society of Agricultural Statistics held at Rajkot, Gujarat on December 6-8, 1997.

Statistics has emerged as a mature discipline during the present century. It has its own foundations and philosophy as well as its techniques. It pervades essentially all aspects of human endeavor. In the agricultural sciences, statistics has played a vital role in increasing production, developing new seeds, cropping practices, advancing plant genetics, forecasting crop production, animal husbandry, agricultural marketing and in many others. The field of agricultural statistics comprises the areas of statistical sciences, which apply directly to the interpretation of results in agricultural sciences. The presentation here is non-technical and speculative. The object of the paper is to appraise the agricultural worker with the newly developed techniques in statistics, which seem highly promising for the advancement of agricultural statistics. The discussion includes some of these techniques and a very brief discourse on each of them. The references at the end of the paper will provide further material for the interested reader.

2. *Twentieth Century Statistics and Fisher's Influence*

Sir Ronald A. Fisher developed a vast amount of statistical knowledge while he was associated with Rothamstead Agricultural Experiment Station. He developed the theory of design of experiment, analysis of data through analysis of variance technique and various related topics. His fundamental contributions of using randomization and replication in experimentation not only revolutionized agricultural research but also many other scientific disciplines. During the early decades of the twentieth century, he laid the foundation of estimation and introduced the concepts of sufficiency and efficiency in statistics. The method of maximum likelihood of obtaining estimates was fully developed through his efforts. He gave the concept of Fisher information and many other notions. Essentially, it can be said that the twentieth century was inaugurated as Fisherian. Pre Fisherian statistics has been heavily dependent on the concept of correlation, regression and the moment method of estimation which may be regarded only as a collection of techniques, but not as a science. It was Fisher who gave statistics its coherence as a scientific discipline in 1925, see Efron [3].

3. *Agricultural Sciences*

From its early roots in human civilization, agriculture started as a technology but now has been transformed into a well-developed science. In modern times, advances in agricultural methods have been accomplished through scientific experimentation and their application to practice by the individual agriculturist. In most countries, agriculture and associated food production industries constitute a huge portion of their economic activity. In agricultural

enterprises, we include sowing, harvesting, and marketing of all crops, dairy cattle, poultry, sheep and fisheries, fruit cultivation, animal and plant husbandry, genetic engineering, agricultural engineering, and biotechnology. There are huge data banks and agricultural censuses, etc., in most countries on agriculture.

The field of agricultural statistics has developed in the past few decades and it deals with the growing amount of information inherent in the agricultural sciences. There are hundreds of books and journals in the field.

In the twentieth century, great strides were made in the advancement of agricultural techniques. The contributions of agricultural statistical methods to this development are unquestionably immense. In the development of agricultural statistics as a mature field, the leadership of the Government of India, Indian Agricultural Statistics Research Institute, New Delhi is no doubt paramount. Fisher developed many leading statistical techniques when he worked on agricultural problems at the Rothamstead Agricultural Experimental Station in England and therefore agricultural research initiated many theoretical advances in statistical sciences.

4. Statistical Century

From the early decades of the twentieth century where the bulk of statistics consisted of a collection of techniques dealing with frequency curves, correlation and regression analysis, and statistics developed as a mature discipline with its philosophy and techniques. Contribution of Sir Ronald A. Fisher to the development of statistics as a mature discipline is immense. He also helped in the applications of the techniques to agricultural sciences. Not only statistics now is applied to various sciences but also a large number of new disciplines have evolved as a result of the development of their statistical aspects. Such influence is evident from the fields of the psychometrics, biometrics, chemometrics, envirometrics, polimetrics, econometrics, and so on. Fields of educational statistics, biostatistics, geostatistics, medical statistics, environmental statistics, and agricultural statistics developed as distinct areas with a body of statistical techniques specific to these disciplines.

Although early statistical techniques dealt with merely correlation coefficient, curve fitting and comparison with coefficients, the twentieth century saw the flowering of statistics as a science of decision making under uncertainty. These developments culminated into the fully developed areas of statistical decision theory, sequential analysis, optimal design of experiments, sampling of human populations, and so on. We discuss next few areas of statistics, which are likely to play a major role in agricultural research and practice in the twenty-first century.

5. *Computer Intensive Methods*

With the advent of the high speed computing, statistical theory in difficult cases has become computational from that of being mathematical. A collection of methods has been developed using computers. Such techniques provide answers in much more general settings than available previously and sometimes answer questions, which could not be answered theoretically or mathematically. The elegance of mathematical proofs has been relegated to empirical verification of some of these procedures. The computer intensive methods play an important role in statistical sciences. We describe a few of the methods, which are likely to play a major role in statistical applications to agricultural sciences.

5.1 Jackknife Method

The jackknife technique helps in reducing bias for statistics where unbiased estimates are not available. For example, it is known that the sample standard deviation based on a random sample calculated by the square root of the quantity which is the sums of the squared deviations from the mean divided by one less than the number of observations in the sample, is not unbiased. Leave-out rules forming the jackknife technique give estimates of the standard deviation with a smaller bias. The techniques consist of leaving, say one observation at a time, and calculating the standard deviation from the remaining observations. Averaging then all the standard deviations calculated gives an estimate that has much smaller bias. This technique has been mathematically justified and extension of the rules based on more general leave-out rules is available.

5.2 Bootstrap Procedure

Bradley Efron has developed the bootstrap procedure as a resampling procedure. It provides estimates of standard errors specifically where they are not easily obtainable. Confidence interval estimates or estimates of quantities of the population as well as other quantities of interest such as those of correlation coefficient, regression coefficients can be numerically obtained using bootstrap methods. The procedure involves taking repeated samples from the empirical distribution functions of the obtained sample. The statistics of interest is calculated from each sample. Lots of such samples—in thousands, are taken; the statistic is computed and thus provides the empirical distribution function of the statistic. Quantiles, standard errors and confidence intervals, etc. then can be obtained. In bootstrap, very general assumptions of the underlying distributions are made. These procedures have been extensively employed in

survey sampling, density estimation, time series analysis, and so on. For a comprehensive recent survey, see Efron and Tibshirani [5].

5.3 Meta Analysis

Meta Analysis deals with techniques that combine analyses from all relevant individual studies into a single statistical analysis with an overall estimate and confidence interval for effect size. The studies include both significant and non-significant results. Great statistical power can be achieved through Meta analysis. Such studies are usually collected from review of literature. There are questions of publication bias in Meta analysis as unpublished studies, by definition, are not included. For a recent discussion, see Givens, Smith and Tweedie [8].

6. Functional Data Analysis

When the datum is a curve or time series and several such data are available for study, we have a problem in functional data analysis. Essentially the datum is a function. Problems of comparing growth curves belong to this field. In agricultural statistics, problems of functional data analysis abound and the twenty-first century will see an expanded use of these techniques. For a recent survey, see Silverman [14].

7. Spatial Statistics

Many applications in agriculture and forestry require data analysis of spatial objects. Satellite data provide many rich examples of the applications of spatial statistics in agriculture. For crop predictions, areas under a given crop, forests, etc., satellite observations have given a new way of understanding agricultural data in the twentieth century. It will play a vital role in the future. For a recent discussion of spatial data analysis, see Cressie [2].

8. Neural Networks

Neural Networks originally were intended as abstract models of the brain. In statistics, the basic structure of the models is used for prediction as well as for understanding the structure of the process to which the models are applied. These models are especially useful in dynamical control theory, artificial intelligence, and machine learning. In many cases of manufacturing, robotics, and engineering sciences, the use of neural networks has made significant contributions and it is expected that they will play a vital role in agricultural science in the future. For a recent discussion, see Cherkassky *et al.* [1].

9. *Markov Chain Monte Carlo (MCMC)*

Using the Markov Chains Monte Carlo procedure is used for integration. Baysians use the Markov Chain Monte Carlo (MCMC) method for integrating over posterior distributions of parameters given the data. Ordinary Monte Carlo methods for integration require taking samples from the required distribution and using sample averages to approximate expected values. Similarly Markov Chain Monte Carlo integration takes samples from Markov chains. Gibbs sampling is a special case of the MCMC method. Early development of MCMC methods is given by Metropolis *et al.* [10] and Hastings [9]. For a general discussion of the MCMC methods, see Gilkes, *et al.* [7].

10. *Image Analysis*

In agricultural settings, especially with the availability of satellite data, it has become important to analyze digital images of interest. The Baysian image analysis is concerned with the image as numerical data generated by a statistical model involving both the random component and a systematic component. Utilizing Baysian techniques, the likelihood combined with the prior distribution on the true scene description, gives inferences about the scene based on the recorded image data. Calculations of image models can be facilitated by the use of Markov Chain Monte Carlo methods. For satellite image analysis, see the recent study by Wilson and Green [16].

ACKNOWLEDGEMENTS

I am grateful to Bal B.P.S. Goel and Sneh Goel for suggesting improvements to an earlier version of the paper. To Prem Goel, I am thankful for suggesting a few key references.

REFERENCES

[1] Cherkassky, V., Friedman, J.H. and Wechsler, H., (editors) 1994. *From statistics to neural networks: Theory and pattern recognition applications,* Springer-Verlag, New York.

[2] Cressie, Noel, A.C., 1993. *Statistics for spatial data,* John Wiley and Sons, New York.

[3] Efron, B., 1997. R.A. Fisher in the 21st Century, Technical Report No.194, Stanford University, Division of Biostatistics, Stanford, CA, 1-48.

[4] Efron, B. and Gous, Alan, 1997. Baysian and Frequentist Model Selection, Technical Report, Stanford University, Department of Statistics, Stanford, CA, 1-28.

[5] Efron, B. and Tibshirani, R.J., 1993. *An introduction to bootstrap*, Chapman and Hall, New York.

[6] Gelman, A., J.B., Carkun, H. Stern and D.B. Rubin, 1995. *Baysian data analysis*, Chapman and Hall, New York.

[7] Gilkes, W.R., S. Richardson, and D.J. Spiegelhalter, 1996. *Markov chain Monte Carlo in practice: Interdisciplinary practice*, Chapman and Hall, New York.

[8] Givens, Geof H., Smith D.D., and Tweedie, R.L., 1997. Publication bias in Meta Analysis: A Bayesian Data-Augmentation Approach to account for issues exemplified in the passive smoking debate, *Statistical Science*, **12**, 221-240.

[9] Hastings, W.K., 1970. Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 97-109.

[10] Metropolis, N., Rosenbluth, A.W. Rosenbluth, M.N., Teller, A.H. and Teller, E., 1953. Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087-1091.

[11] Neal, Radford M., 1996. *Bayesian learning for neural networks*, Springer, New York.

[12] Ripley, Brian D., 1996. *Pattern recognition and neural networks*, Cambridge University Press, Cambridge.

[13] Rustagi, J.S., 1990. Computer intensive methods in agricultural statistics, *Jour. Ind. Soc. Agri. Statist.*, **42(3)** 259-276.

[14] Silverman, B., 1997. *Functional data analysis*, Springer-Verlag, New York.

[15] Wahba, G., 1990. *Spline models for observational data*, Society of Industrial and Applied Mathematics, Philadelphia.

[16] Wilson, J.D., and Green, P.J., 1993. A Baysian Analysis if remotely sensed data, using Hierarchical Model. Research Report S-93-02, School of Mathematics, University of Bristol, England.