Report on the Session on Statistical Computing held during 47th Annual Conference of Indian Society of Agricultural Statistics at S.V.Agricultural College, Tirupati on 18th December, 1993.

The ISAS organised a session on Statistical Computing during its annual Conference to discuss the use of computer, intensive methods for research in Agricultural Statistics, instead of simply focussing attention on application of Computer and its related problems in usual data analysis. The sub-topics of interest identified thus were Bootstrapping, Jackknife, Monte- carlo Simulation, Computer Modelling and Information technology. The Session had as

Chairman : Prof. T.V. Hanurav
ISI, Hyderabad

Convenor : Dr. V.K.Bhatia
IASRI, New Delhi

At the outset, Secretary of the ISAS welcomed the Chairman and highlighted the importance of this Session particularly in the light of emergence of newer Computer-intensive methods.

After a brief opening remarks about the Session, Chairman invited speakers to present their papers.

Dr. U.M. Bhaskar Rao of CRIDA, Hyderabad presented the paper entitled 'Use of Generalised Linear Models in Analysing Dryland Research Data'. He emphasised that the computer interactive approaches should be used to deal with the situations where the usual assumptions of constancy of variance, normality, additivity are violated. He also pointed out that the Software GLIM (Generalised Linear Interactive Modelling) can be used in this direction more efficiently. The usefulness of this technique is highlighted by actual applying it to the uniformity trial data on Sorghum collected under dryland conditions. He further concluded indicating the other use of GLIM in arriving at more realistic models dealing with crop planning strategies under dryland conditions.

Dr. V.K. Bhatia of IASRI, New Delhi through his paper entitled 'Computer Intensive Method for studying Standard Deviation and Confidence Interval of Genetic Parameters' focussed attention on use of Bootstrap technique in obtaining standard error and confidence interval of one of the important parameter, the heritability co-efficient. After pointing out in brief the theoretical background of the bootstrap technique he showed its usefulness by actual applying it to the simulated and real data. From the results obtained, he

concluded that this technique is very reliable analytical procedure for obtaining the standard error and confidence interval of the complex parameters whose explicit expressions for second moments, variance on confidence intervals etc. are either not available or very difficult to obtain. He was also of the opinion that bootstrap technology may go a long way in solving many problems particularly in Statistical Genetics where at times things get struck down for the lack of theory.

Sh. R.L. Sapra of NBPGR, New Delhi presented his paper entitled 'Entity Relationship Modelling for Germplasm Information System'. While discussing the information needs of an organisation, he highlighted the importance of computer based systems approach. He pointed out that with proper data base and system approach, one can develop a suitable Germplasm Information system for storing and disseminating the information requirements of various organisation of Plant Genetic Resources in India.

Sh. S.C. Tewary of GBPAU, Pantnagar spoke on the 'Management System for Computing the ANOVA'. He explained, how he had developed a software for analysing agricultural field experimental data based on manipulating various sum of squares of ANOVA table. This, he had achieved by basically bit manipulation and mapping based on empirical relations. He also stressed that the applicability of this software is much wider in comparison to the existing softwares like SPSS, BMDP etc.

After these presentations, chairman invited Prof. Prem Narain for his observations on the Computer-Intensive Methods. Prof. Prem Narain emphasised that Computational aspects of statistical theory is a very important component right from the Fisher times. He pointed out that how with the help of only small computing machine, Fisher could come up with the concepts of Randomisation, Replication and Local control. He further focussed the attention that although the concept of correlated observations was identified by Papadikas way back in 1937 but in the present scenario only, with the advent of fast computers and revolution in the field of electronic computation, a new look is given to the statistical theory. He highlighted the usefulness of the various computer-intensive methods such as jackknife, bootstrap, non-parametric regression, generalised additive models etc. He also cited an example of non-parametric regression between two variables popularly known as LOESS. In this method, no model is assumed over the entire range of the independent variable but instead a series of local regression curves for different values of the target point are fitted. The process is repeated for all possible target points to obtain a non-parametric regression. Finally, he concluded by highlighting the importance of information technology in the present day need for carrying out research in the field of Agricultural Statistics.

In his closing remarks, chairman highlighted the importance of examination and computational aspects of the data. He had put a lot of emphasis on the rigorous examination of the data to derive as much information as possible. In this regard, he recalled the era of Prof. Haldane and Prof. Fisher where they used to carry out the entire calculation work themselves with their own hands. He also emphasised that although newer computer-intensive methods are very useful in achieving much information but these should not be used as a matter of practice and moreso at the cost of theory. He was of the opinion that howsoever complex statistics may be, firstly efforts should be made to obtain theoretical solution by following theoretical principles and if theory fails then only one should go for computer intensive methods. He reiterated that one should look into the use of computer intensive methods and statistical packages rather than their misuses. In other words, we should not translate our problem as per specification of the package but try to use it in carrying out sophisticated analysis. He warned that over dependence on these packages and methods may not prove to be good for the statistical theory.

Recommendations

i) Computer-intensive methods should only be used where theory is unable to provide results.

ii) Over dependence on statistical packages should be avoided and these may be used as tools only.

iii) Examination and computational aspects of data may be carried out more rigorously.

iv) The use of bootstrap in the area of statistical genetics needs be explored.

v) Information technology and software development should get due attention.

1. Use of Generalized Linear Models in Analysing Dryland and Research Data

M. Narayana Reddy* and U.M.B. Rao*

The statistical problems of both design and analysis of dryland experiments need more attention of the statisticians as the soil exhibits uneven and high variability on crop yields under different treatments. This results in many times high standard errors when analysed through ANOVA with usual assumptions. Generalized Linear Model (GLM) introduced by Nelder and Weddenburn (1972) is useful in investigating the new approaches in statistical analysis through computer interactive approach in which the usual assumptions such as constancy

* Central Research Institute for Dryland Agriculture, Hyderabad.

of variance, normality, additivity are no longer a requirement. The software called Generalized Linear Interactive Modelling (GLIM) which was initially developed by Baker, Clarke and Nelder enables to specify model for the data, to find the best subsets from a class of models and to examine further the implications of fitting such models. A particular GLM can be identified by (i) specifying the error distribution, (ii) the form of the linear predictor and (iii) the function linking the mean of the linear predictor. Many problems related to the analysis of agricultural research data such as modelling for (i) fertilizer application (ii) plant density (iii) pest disease control and (iv) rainfall etc. can be carried out by using GLM algorithm. A detailed analysis was carried out for examining the adequacy in fitting the two dimensional Smith's variance law with the five years uniformity trial data on sorghum collected under dryland conditions using GLM and non-linear procedures. These results are compared with the results obtained through linear regression analysis. The residual plots indicate GLM and non-linear models to be more adequate compared to linear models. The estimates obtained through GLM are more acceptable due to the statistical properties of their estimators compared to non-linear estimators. Relations explaining the variability between plant density and crop yield were also examined through GLM procedure. It was observed that the quadratic polynomial with gamma distribution of errors is performing better for the two sets of data that was examined. There is a need for further work in this aspect. Rainfall analysis is an important area where GLIM is useful to arrive at the more realistic models, particularly to arrive at the crop planning strategies under dryland conditions.

2. Computer Intensive Method for Studying Standard Deviation and Confidence Interval of Genetic Parameters.

V.K. Bhatia*, J. Jayasanker* and S.D. Wahi*

The knowledge of genetic parameters namely heritability, repeatability and genetic correlation are very much useful in formulating selection strategies for genetic improvement. Although a number of good estimators are available for estimation of these parameters but there is a great scarcity of trustworthy estimators for their precision and reliability. The available estimators of precision are generally based on a number of approximations and assumptions. This is more so true in the case of confidence interval estimation. These methods are based on the assumption that the observations are normally distributed. Very often, this assumption does not hold and it is not known that how sensitive these methods are to the normality assumption. In recent years some more robust and computer intensive methods based on resampling techniques are developed.

* Indian Agricultural Statistics Research Institute, New Delhi

The most common of these methods are the Jackknife and the Bootstrap methods.

Different research workers in the past used the Jackknife method for estimation of the sample variance of some genetical components. In some cases, it is shown that the transformation of the data may give very good results and in other situations it may be impossible to find a suitable transformation. Though different researchers tried to give Jackknife and bootstrap variance estimates of genotype correlation but subsequently it was found difficult to give reasonable interpretation of its sampling variance. Therefore, it is desired to estimate a confidence interval for these parameters. With this in mind the present paper following Aastveit (1990) is focussed on obtaining an estimate of the standard deviation and also a confidence interval for the parameter heritability. Different resampling schemes exist for both the Jackknife and Bootstrap procedures. The present study, however, deals with only Bootstrap procedure. The data sets with varying values of heritability estimates are taken into account and results so obtained are compared with that obtained under normality assumptions.

3. Entity Relationship Modelling for Germplasm Information System

R.L. Sapra*

Entity Relationship Modelling is an information engineering technique used to develop high quality data model in any organisation and is being globally adopted for defining the information need of an organisation with greater integrity of the data. System analysts use this technique extensively for improving system quality and software productivity. The technique is also being used for the development of Germplasm Information systems in some of the genebanks operating in the world for the management and effective use of valuable germplasm resources. The present paper describes, in brief, some of the concepts and definitions associated with the Entity Relationship Modelling with special reference to genebank information. The paper also describes the tentative relational data model developed at the Germplasm Resources Information Network (GRIN) unit of USDA, Beltsville, USA for defining the information requirements of National Bureau of Plant Genetic Resources (NBPGR), New Delhi. The developed model which contain 23 entities/tables with varying number of attributes is in the process of being implemented on a 486 EISA server using the ORACLE ver 7 Relational Database Management System (RDBMS).

* National Bureau of Plant Genetic Resources, New Delhi

4. Management System for Computing the ANOVA

V.K. Srivastava* and S.C. Tewary*

In almost all fields of studies in agriculture, biology, engineering, social science etc. generally the main purpose of conducting an experiment is to make comparisons among various treatments. For example, in case of agricultural research one conducts an experiment to compare grain yields of say three rice varieties under three management practices and five applications of nitrogen doses. For this purpose a split-split-plot design can be used with nitrogen as main plot factor, management practices as sub- plot factor and variety as sub-sub-plot factor. Consider another example from food processing engineering to study preservation of khoa under three preservatives, two temperature and two storage time. Thus, here the treatment includes three levels of a preservatives, two temperatures and the two storage time. So we have a three-factor experiment in completely randomized design. To analyze the data arising from the experiments of the type described above the algebraic analysis of variance (ANOVA) technique as given by Fisher is used in which hypotheses about the various components are tested. Various software packages like SPSS, BMD, etc. deal with narrow domain and can be used only for the commonly used designs. If the user is confronted with a new design he has to write a new program as the above packages are not flexible.

In the present paper a state-of-the-art technology has been developed for computation of ANOVA table based on the design as per user's specification i.e., by management of various sums of squares. The salient features of this package are :

- Executable version requires 14k in MSDOS.
- It has portability at different hardware plateforms.
- It runs without overlays and OS overheads.

For data based on 'n' number of factors, the combinations of sums of squares are $2^n - 1$. The ANOVA table of the design under consideration is computed by management from the sums of squares of these $2^n - 1$ combinations. The management takes effect by mapping based on empirical relation.

* G.B.P.U. Agriculture and Technology, Pantnagar