Indian J. pure appl. Math., **26**(6) : 521-529, June 1995

STATISTICS AND ITS APPLICATIONS TO AGRICULTURE AND GENETICS

P. Narain

Indian Agricultural Research Institute, Pusa, New Delhi 110 012

Statistics has important and interesting interfaces with agriculture including genetics and breeding of crop plants and animals. The paper discusses some of them with concrete applications to problems in agriculture and genetics. In particular, measuring change in agricultural surveys, partial diallel crosses useful in plant breeding, stochastic processes in population genetics and statistical prediction in animal breeding are discussed.

1. Introduction

The subject of statistics as we know deals with random phenomenon and uncertainity. Its foundation is based on mathematics and probability and that part of the subject has acquired the name of 'mathematical statistics'. As an applied science, however, it has grown into a body of knowledge which deals with measuring and minimising uncertainty surrounding the data generated in a given discipline like biology, agriculture, engineering etc. In fact, it is at the interface of statistical methods and the discipline of application that statistics really exists in the form of a science of the meaning and use of data. This makes statistical science as an inter- and cross-disciplinary research activity. It is this characteristic of the subject of statistics which Prof. Mahalanobis emphasised way back in thirties and to this end founded the Indian Statistical Institute, Calcutta where besides statistics divisions, other units like anthropometry and human genetics, agricultural science and physical and earth sciences division etc. were set up to provide ways and means of interfacing statistics with disciplines of application. Prof. Mahalanobis was indeed a visionary who could foresee that statistics could only flourish if data required for this purpose are generated in the Institute itself. A developed country like USA has only recently realised this feature of statistics as can be seen from a report of Olkin and Sacks[14], Co-chairmen of a Panel of the Institute of Mathematical Statistics and the consequent setting up of an Institute of Statistical Sciences in North Carolina, USA.

I was fortunate indeed to have worked at the interface of statistics with agriculture and genetics. At the Indian Agricultural Statistics Research Institute (IASRI) where I spent almost thirty years of my career and where I was Director for over ten years, I contributed to the development of such an interface. In fact, right from the very inception of the IASRI in 1930 as well as the Indian Society of Agricultural Statistics (ISAS) in 1947, both located at New Delhi, this interface between statistics and agriculture was emphasised which ultimately led to the emergence of the discipline of 'agricultural statistics'. Bearing this in mind I will briefly discuss some of the 'concrete' applications of statistics to agriculture and genetics.

2. Measuring Change in Agricultural Surveys

In any development process, efforts are made to bring about desirable changes and their measurement is of prime importance. Sample surveys, as an approach to data collection and for deriving meaningful conclusions within limited resources of budget and time constraints, are then the obvious choice. An important feature of the agricultural surveys in India is the availability of hierarchical structure of the geographical units like villages, households or cultivators within a village and fields or animals within a household or with a cultivator. This provides natural choices for sampling units at different stages of selection. Seasonality, as for instance in milk production, is another significant feature in these surveys which requires estimating changes over the years as well as changes within years. Repeating the survey on several occasions is the common approach for estimating these changes. An important issue in this context is the estimation of changes in populations integrating simultaneous study of several similar characters on the basis of commoness of problems and data collection approach. We discuss here briefly one such problem of estimating the production of various livestock products like milk, eggs, wool and meat through continuous surveys in which one product is estimated with higher precision in one year while in other years changes are estimated on the basis of smaller sample. The details can be seen in Narain *et al.*[12].

The sampling plan is that of a stratified multistage sampling with geographical strata consisting of tehsils/talukas, clusters of two adjacent villages as primary sampling units (psu's) and households as second stage units (ssu's). A year consists of three seasons viz. summer, rainy and winter. With a rotation plan for three years, successive sampling was used over the nine seasons (occasions) of the three years in the selection of psu's. A sample of m psu's was matched over all the nine seasons (I set). A sample of $(n - m)$ psu's was matched over the same seasons in different years (II set). In a specific year, this sample of n psu's was supplemented by $(n' - n)$ psu's in each season (III set).

If we take a specific character like milk, the parameters of interest are seasonal and yearly averages and changes of number of animals in milk/milch animals, total milk production and average milk yield per day per animal in milk. Let x denote the character as the number of animals in milk, $\bar{x}_{ij(t)}$ $(t = 1, 2, 3)$ be the average number of animals in milk per psu for the t-th set in the jth season of the ith year. Similarly we can define for number of milch animals and milk production. Further $\bar{x}_{ij(1+2+3)}$ denotes the average based on total data pooled over sets I, II and III defined above.

With milk as the main character, there are years of two types — year I in which milk is the main character and years II and III in which it is the secondary character when main character could be eggs/wool/meat. The Minimum Variance Linear

Unbiased Estimator (MVLUE) of $\hat{\bar{X}}_{ij}$, the average number of animals in milk per psu may be obtained as

$$\hat{\bar{X}}_{ij} = \sum_{j=1}^{3} \sum_{t=1}^{3} a_{1\,j(t)} \left(\bar{x}_{1\,j(1+2+3)} - \bar{x}_{1j(t)} \right)$$

$$+ \sum_{i'=2}^{3} \sum_{j=1}^{3} \left(\bar{x}_{i'\,j(1)} - \bar{x}_{i'\,j(2)} \right) + \bar{x}_{ij\,(t')}$$

where $t' = 3$ for $i = 1$ and $t' = 2$ for $i = 2$ and 3. Here a' s may be optimised by minimising the variance of the estimator. In view of Eckler's result[1], this form of the linear estimator will lead to MVLUE. This estimator has got 15 coefficients of a's and may be written as

$$\hat{\bar{X}}_{ij} = \sum_{i=1}^{15} a_i z_i + \bar{x}_{ij\,(t')}$$

where z_i's are zero functions with their expectations as zero. This leads to equation of the form

$$\underset{\sim}{P} A = B$$

where $\underset{\sim}{P}$ is the variance-covariance matrix of

$$z = (z_1, \ldots, z_{15}),$$

$\boldsymbol{A}$ is the vector of a's and $\boldsymbol{B}$ is the vector for covariances between z_i's and $\bar{x}_{ij(t)}$. With these MVLUE estimators of seasonal averages, any linear combinations of these parameters can be estimated by the same linear function of the corresponding estimators.

Although this procedure ensures minimum variance linear unbiased estimation of various seasonal and yearly averages and their changes, the procedure is somewhat complex, requiring solution of a large number of equations. The simplicity of recurrence relationships of MVLUE estimators, normally available in successive sampling under specific correlation models given by Yates[18], Patterson[15] and Tikkiwal[17], is not available here in view of the seasonality of the characters. The correlations between seasons of the same year are expected to decline with time but those between the same seasons of consecutive years are likely to be high again. A somewhat approximate approach may be to resort to various difference estimators or some linear estimators with lesser number of constants. One such estimator for the average number of animals in milk per psu for the jth season of first year was considered as

$$\hat{\bar{X}}_{ij} = a_{1j} \left(\bar{x}_{11(2+3)} - \bar{x}_{11(1)} \right) + b_{1j} \left(\bar{x}_{12(2+3)} - \bar{x}_{12(1)} \right)$$

$$+ c_{1j} \left(\bar{x}_{13(2+3)} - \bar{x}_{13(1)} \right) + \bar{x}_{1j(2+3)}, \quad (j = 1, 2, 3)$$

where a, b and c were obtained by minimising the variance of $\hat{\bar{X}}_{ij}$ and $\bar{X}_{1j(2+3)}$ denotes the average based on sets II and III combined. Evidently, this estimator was not MVLUE but it performed reasonably well.

3. Partial Diallel Crosses Based on Extended Triangular Association Scheme

Often a plant breeder is required to evaluate the general combining ability (gca) and specific combining ability (sca) involving a large number of inbred lines for choosing the best cross amongst them. Sometimes, he is required to estimate the genetic components of variance and covariance for various economic characters with a view to exploiting them by suitable breeding methods. For this purpose 'diallel cross' technique, in which all possible single crosses among a group of inbred lines are raised, is used. With the exclusion of reciprocal crosses and parental inbreds, there are $N = n(n-1)/2$ possible single crosses among a set of n lines which are to be tested in a suitably replicated randomised design. This number increases rapidly with increase in n. With facilities available for testing only a limited number of crosses, a diallel cross may therefore be possible only when n is relatively small. However, if a small number of lines only are included, the estimates of the variances of the gca and sca among the whole population of potentially available lines is subject to large sampling errors and many potentially high yielding lines may be left completely untested. It is, therefore, necessary to have a large number of inbred lines but raise only a sample of all possible crosses amongst them. Such a diallel cross is known as 'partial diallel cross' (PDC). Several workers such as Kempthorne, Curnow, Narain, Arya and others have discussed various statistical designs for performing the PDC's. Often the structure of a partially balanced incomplete block (PBIB) design is made use of. This is because of the one-to-one correspondence between the complete diallel cross (CDC) and balanced incomplete block (BIB) design with 2-plot blocks. Similarly, there is a one-to-one correspondence between PDC and PBIB designs with two plots per block and two associate classes or three associate classes or in general m associate classes. In this case, however, the property of balance possible with CDC is disturbed and we may have different variances for different comparisons. The efficiency of the PDC then depends on the average of the variances over all the comparisons. We discuss briefly the PDC based on extended triangular association scheme (Narain *et al.*[11]).

Let the number of parents n be of the form $p(p-1)(p-2)/6$ where p is an integer greater than 3. Now denote a parent by a triplet abc, where a takes any value from 3 to p, b takes values from 2 to $(a-1)$ and c takes values from 1 to $(b-1)$. All the parents can then be numbered off into $(p-2)$ different triangles $T_1, T_2, ..., T_{(p-4)}, T_{(p-3)}$ and $T_{(p-2)}$ of orders $(p-2) \times (p-2)$, $(p-3) \times (p-3)$ ... 2×2 and 1×1 respectively, the number of parents (triplets) in the ith triangle T_i, therefore being $(p-i)(p-i-1)/2$. Now three different types of PDCs known as Extended Triangular Design (ETD) can be constructed :

Design I

We sample all the crosses of the type *abc* × *def*, where *a*, *b*, *c*, *d*, *e* and *f* are all distinct. The number of times the parent p $(p - 1)$ $(p - 2)$ of the triangle T_1 is involved in crossing with other parents is then

$$s_1 = \sum_{i=4}^{(p-2)} (p-i)(p-i-1)/2 = (p-3)(p-4)(p-5).$$

The same is true for every other parent. The resulting sample would then consist of $(ns_1)/2$ crosses.

This procedure corresponds to picking up the third associates of each treatment in the extended form of triangular association scheme given by John[5] and pairing the treatment with each member of the associate class. The number of third associates would be s_1.

Design II

We sample all the crosses of the type *abc* × *def* where one of the letters is common resulting in three categories of crosses. The number of times each parent is involved in crosses with other parents becomes

$$s_2 = 3(p - 3)(p - 4)/2$$

which happens to be the number of second associates of each treatment in the corresponding design.

Design III

We sample all the crosses of the type *abc* × *def* where two of the letters are in common leading to three categories of crosses. The number of times each parent is involved in crosses with other parents then becomes

$$s_3 = 3(p - 3)$$

which happens to be number of first associates of each treatment in the design.

The analysis of PDC constructed above follows the pattern of the analysis of three-associate PBIB design given in Rao[16].

The average variances of the differences between the *gca* effects of any two parents happen to depend on p. It decreases with increase in the value of s_1 (or s_2 or s_3). The efficiency of the ETD designs vis-a-vis Circulant Design (CD) of Kempthorne and Curnow[6], for the same number of crosses sampled, is always greater than one. The Design I is found to be much more efficient than either Design II or Design III.

4. Stochastic Processes in Population Genetics

Perhaps the most important and advanced application of statistical science to genetics is in relation to the development of mathematical theory of population genetics. Fisher, Haldane and Wright contributed significantly by developing deterministic and stochastic models in this context. However, application of stochastic processes, particularly the diffusion process to population genetics, indicating the effect of finite population was greatly advanced by the works of Kimura of Japan and Crow of USA. This development initiated vigorous efforts by eminent mathematicians and statisticians to understand the process of gene substitution not only in classical terms but also at the molecular level.

When we study genetic differences between individuals in an infinitely large random mating population, the most important principle is Hardy-Weinberg law on the constancy of gene and genotypic frequencies in the absence of directed forces to change the frequencies of genes in a particular direction. Selective processes due to differential fertility of parent modify the Hardy-Weinberg law. In this connection, one of the important principles having bearing on natural selection is that of Fisher's 'Fundamental Theorem of Natural Selection' which states that the rate of increase in the average fitness of the population is equal to the additive genetic variance in fitness of that population. Random changes in gene frequency due to finite population size results in random fluctuations of gene frequency in a population over time because of random sampling of gametes necessary to form a new generation. This phenomenon has been given the name 'random genetic drift'.

Mathematically, the stochastic process of gene frequency change can be approximated as a diffusion process with the random variable × representing the frequency of a gene A with time parameter (t) changing continuously. At a particular given time, we have now a 'gene frequency distribution' with density function $f(x, t)$. It is possible to show, by methods often used in physics, that this density function satisfies Kolmogorov forward equation

$$\frac{\partial f(x,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}[v(x) f(x,t)] - \frac{\partial}{\partial x}[m(x) f(x,t)]$$

with $m(x)$ and $v(x)$ as given instantaneous drift and diffusion coefficients respectively. These represent respectively the expected mean change as well as the variance of change in gene frequency. By taking particular values of these various parameters and solving the resulting partial differential equations, it is possible to arrive at an analytical solution for the gene frequency distribution under various situations. Besides, one can also determine the 'probability of fixation' of a mutant, being favoured during selection by solving the Kolmogorov backward equation

$$\frac{\partial f(q,x;t)}{\partial t} = \frac{v(q)}{2}\frac{\partial^2 f(q,x;t)}{\partial q^2} + m(q)\frac{\partial f(q,x;t)}{\partial q}$$

where q, the initial gene frequency at time $t = 0$, is now a random variable with frequency x as fixed and the process is considered retrospectively by reversing the time sequence. When $x = 1$, $f(q, 1; t)$ is denoted by $u(q, t)$, the probability of fixation of a gene by time t.

In considering the population dynamics of mutant substitution, we need, in addition to the probability of gene fixation, the 'average length of time involved for each gene substitution'. For getting quantitative estimates of time to fixation, a general theory based on diffusion approximation as well as Markov chain methodology was developed by Narain[7,8]. This involves conditioning the diffusion process or the Markov chain for the contingency of eventual fixation of the gene. In the former case it leads to conditioned forward as well as backward diffusion equations with modified drift coefficients but with the same diffusion coefficients as in the case of the unconditional process. Using the backward form of the conditioned diffusion equation, one can then develop differential equations for obtaining the various moments of the distribution of time until fixation of a gene. For instance, in the case of pure random drift $m(q) = 0$, $v(q) = q(1 - q)/2N_e$ where N_e denotes the 'effective' population size which is approximately equal to the number of breeding individuals in one generation and is usually smaller than N, the population size, due to the distribution of progeny number per individual deviating from Poisson distribution with mean 2. Then the average time until fixation is found to be

$$M_c(q) = -4N_e \frac{(1-q)}{q} \log_e (1-q).$$

5. Statistical Prediction in Animal Breeding

Statistics is concerned with the development of scientific methods of induction for prediction from 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, Pearson and Wald. In particular, Fisher[2], while talking about the estimation stated that the objective of the statistical method is 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 a model, say,

$$Y = a(1 - \exp(-k(x - b))).$$

Then we take the observed yields 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 error e is assumed to be 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 statistic 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.

It seems the whole issue can be given a unified treatment by developing a general prediction theory (Harville[3]). 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 second moments denoted by

$$\mu_x = E(x),\ \ \mu_y = E(y),\ \ \sigma_x^2 = \mathrm{Var}(x),\ \ \sigma_{yx} = \mathrm{Cov}\,(y, x)\ \text{ and }\ V_y = \mathrm{var}\,(y).$$

We assume that $\boldsymbol{\mu}_y$ belongs to a known vector space and that μ_x is a known linear combination of the elements of $\boldsymbol{\mu}_y$. This means $\boldsymbol{\mu}_y = \underset{\sim}{X}\beta$ and $\mu_x = \boldsymbol{\lambda}^T \boldsymbol{\beta}$ where β is a $p \times 1$ vector of unknown parameters, $\underset{\sim}{X}$ is a $n \times p$ known matrix of rank p^* and λ is a $p \times 1$ known vector that is expressible as $\boldsymbol{\lambda} = \underset{\sim}{X}^T \boldsymbol{k}$ for some vector $\boldsymbol{k}$. The quantity σ_x^2 and the elements of σ_{yx} and $\underset{\sim}{V}_y$ are assumed to be known functions of an unknown parameter vector $\boldsymbol{\theta}$ whose value is restricted to a known set Ω and $\underset{\sim}{V}_y$ is assumed to be non-singular (for all $\boldsymbol{\theta} \in \Omega$). Apart from the linearity of the mean structure, these assumptions mean that σ_x^2, σ_{yx} and $\underset{\sim}{V}_y$ are unrelated to μ_x and $\boldsymbol{\mu}_y$. When we take the special case with $\sigma_x^2 = 0$, x equals $\boldsymbol{\lambda}^T \boldsymbol{\beta}$ with probability one. The problem of predicting the value of x is then equivalent to that of inference about fixed effects $x = \boldsymbol{\lambda}^T \boldsymbol{\beta}$ 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, introduced for the first time by Henderson[4], 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 correlation between the conditional mean of the genetic value, given the records of progeny and the progeny mean 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[9,10,13].

REFERENCES

1. A. R. Eckler, *Ann. Math. Statist.* **26** (1955), 664-85.
2. R. A. Fisher, *Proc. Camb. Phil. Soc.* **22** (1925), 700-25.
3. D. A. Harville, In : *Advances in Statistical Methods for Genetic Improvement of Livestock* (Eds. D. Gianola and K. Hammond), *Springer, New York, 1990, pp. 239-76.*
4. C. R. Henderson, In : *Statistical Genetics and Plant Breeding* (Eds. W. D. Hanson and H. F. Robinson), National Academy of Sciences - National Research Council, Washington, 1963, pp. 141-63.
5. P. W. M. John, *J. Roy. Statist. Soc.* **B28** (1966), 361-65.
6. O. Kempthorne and R. N. Curnow, *Biometrics* **17** (1961), 229-50.
7. P. Narain, *J. Roy. Statist. Soc.* **B36** (1974), 258-66.
8. P. Narain, *J. Genetics* **63** (1977), 49-62.
9. P. Narain, *Biometrics* **41** (1985), 895-907.
10. P. Narain, (1990). *Statistical Genetics*, John Wiley & Sons, New York and Wiley Eastern Limited, New Delhi, 1990 (Reprinted in 1993).
11. P. Narain, C. Subba Rao and A. K. Nigam, *Indian J. Genet. & Plant Breeding* **34**(3) (1974), 309-17.
12. P. Narain, O. P. Kathuria and A. K. Srivastava, *Proc. 46th Session of International Statistical Institute held at Tokyo*, Vol. LII - 2, IP 8.1 (1987), pp. 427-41.
13. P. Narain and A. P. Kaur, *J. British Soc. Animal Production* **58** (1994), 189-96.
14. I. Olkin and J. Sacks. Co-chairs, *Cross-disciplinary researches in the statistical sciences.* A Report of the Panel of the Institute of Mathematical Statistics, USA, National Science Foundation, 1988.
15. H. D. Patterson, *J. Roy. Statist. Soc.* **B12** (1950), 241-55.
16. C. R. Rao, *JASA 42* (1947), 541-61.
17. B. D. Tikkiwal, *Theory of Successive Sampling.* Unpublished Diploma Thesis, ICAR, New Delhi, 1951.
18. F. Yates, (1949). *Sampling Methods for Census and Surveys.* Charles Griffin and Company Ltd., London, 1949.