Jour. Ind. Soc. Ag. Statistics
Vol. XLI, No. 3 (1989), pp. 239-256

BIOMETRY IN FRONTIER SCIENCES*

B. R. MURTY
INSA Senior Scientist, New Delhi

When I received the invitation from Dr. Prem Narain to deliver Dr. Panse Memorial Lecture this year, I considered it an honour and also an opportunity to pay my own humble tribute to this eminent scientist. It is equally gratifying to me to do so under the chairmanship of the illustrious statistician Dr. P. V. Sukhatme who nourished this institution. I wish to recall three events which occurred in rapid succession within a month in 1960 on my return journey from USA visiting Cambridge and Rome and then Delhi. I was fortunate to meet and receive the valuable advice of three distinguished statisticians, which influenced my subsequent research career. When I called on Sir R. A. Fisher in Cambridge, he advised me with paternal affection that my background both in statistics and genetics should be channeled to demonstrate the power of statistics by application to practical problems. Then, I called on Dr. Sukhatme in his office in Rome. He emphasised on the importance of continuous interaction between theory and application from the planning stage of experiments and asked me to see Dr. Panse on arrival in Delhi Dr Panse was kind enough to invite me to his residence and in his characteristic way cautioned me on the quality of data. For me, these three aspects are relevant even today and my gratitude is due to them for their advice. My choice of this topic which involves the three aspects I mentioned now, is determined by the substantial impact of biometry on agriculture and allied sciences and its potential in frontier technology. An early interaction between biometry and the emerging frontier sciences will be mutually beneficial, as all of us share the conviction that "contact with live problems is essential for worthwhile research in statistical metho-

* 'Dr. V. G. Panse Memorial Lecture' delivered at IASRI, New Delhi on 8th November, 1989.

dology." Dr. Mellor, in an earlier Panse Memorial Lecture in 1980, succinctly pointed out, "the intelligence, technical knowledge and the intuitive insights of Dr. Panse in defining the data needs and to his perspicacity and persistence in development and effective operation of major elements of the institutional structure for meeting those needs." This is equally true of the need for such an outlook for the interaction, I mentioned earlier.

While there are several areas of potential application, I wish to concentrate on cases like biotechnology, extension of analogy of biological situations to space science, diagnostic and preventive medical research, remote sensing and environmental monitoring and informatics, with illustrations using available methods, the scope for further research in theory and methodology, and the need for a fundamental research and training wing for the above activity in statistics in our country.

Molecular Biology and Biotechnology in Evolution

The large investments all over the world in biotechnology anticipate major changes in the future of mankind, and will need critical statistical analysis of data. It is established that DNA sequence reflected in the various base composition of different species determine their distinct properties. When we insert alien DNA into this sequence, it changes this sequence and might alter gene regulation and function. The groups of three adjacent nucleotides which code for a specific amino acid are called codons. Therefore, in the analysis of evolution, the substitution processes in homologous DNA sequences are examined in the cell constituents like cytoplasm, mitochondria and others like plasmids in diverse biological groups as plants and animals including man. Substitution of one or more bases by mutation or other processes do change the sequence and codon. Therefore, the estimation of the rates at which such substitutions occur in the paired homologous coding positions in two different species and the divergent times is of interest. Recently, intensified statistical research is in progress on these aspects.

Substitution Rates

Let (1) X and Y species diverged from a common ancestor t years ago, (2) the homologous DNA sequences are of equal length in each, (3) the genes we examine are identical within each member of the species and (4) the four bases A, C, G, T be denoted 1, 2, 3 and 4 for statistical convenience.

Rate of Substitution at Single Site

Considering substitution at a single site, the model for stochastic behaviour of the process is $\{X(t), Y(t), t \geqslant 0\}$ where $X(t)$ and $Y(t)$ denote the nucleotides at a homologous site of the two species at time t after divergence. Thus, $X(0) = Y(0)$ and $X(\cdot)$ and $Y(\cdot)$ evolve independently for $t \geqslant 0$. Then

$$f_{ij}(t) = P\,[X(t) = i,\ Y(t) = j \mid X(0) = Y(0)]$$

$$= \sum_{k=1}^{4} \pi_k\, P^x_{ki}(t)\, P^y_{kj}(t) \qquad (1)$$

where

$$P^x_{ij}(t) = P\,[x(t) = j \mid x(0) = i]$$

and $P^y_{ij}(t)$ is analogous, and π_k is the probability that the base in the common ancestor was k i.e. $\pi_k = P\,[X(0) = Y(0) = k]$.

New, we can describe the transition function P^x_t to model the evolutionary process, where

$$P^x_t = (P^x_{kl}(t)) \text{ and } P^y_t = (P^y_{kj}(t)).$$

As this substitution process is stochastic in nature, we have the well-known continuous time-homogeneous Markov Chain

$$P^r_t = \exp\{Q_r\, t\} = \sum_{n=0}^{\infty} Q^n_r \cdot \frac{t^n}{n!},\ t > 0,\ r = x, y \qquad (2)$$

where $Q_r = (q^r_{ij})$ is the generator P^r_t and is subject to the condition

$$0 \leqslant q^r_{ij} \leqslant \infty \qquad \text{for } i \neq j$$

$$0 \leqslant q^r_i = -\,q^r_{ii} \qquad \text{for } i = j, \text{ and}$$

$$Q^r_1 = 0 \text{ where } 0 = (0, 0, 0, 0)' \text{ and } 1 = (1, 1, 1, 1)$$

q^r_{ij} has an approximation of order $0(h)$ to the probability of a change from base i to j in a small time interval h. Therefore, F_t is symmetric and $F_t = P_t' F_0 P_t$ where $F_t = \{f_{ij}(t)\}$ and

$$F_0 = \text{diag}\,[\pi_1, \pi_2, \pi_3, \pi_4].$$

Thus, we have the generator

$$Q = \lambda \begin{bmatrix} -1 & 1/3 & 1/3 & 1/3 \\ & -1 & 1/3 & 1/3 \\ & & -1 & 1/3 \\ & & & -1 \end{bmatrix} \text{ and } \pi' = [1/4, 1/4, 1/4, 1/4]$$

Therefore, substitution in this model occurs at the points of a Poisson proccess with rate λ.

General Reversible Model

Let us make a stationary assumption that there is a similarity in the evolution process before and after divergence with a distribution frequency of the bases that does not change with time, i.e. $\pi'_Q = 0'$. With the above assumptions, the general reversible model proposed by Tavare (1985) is

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & \bullet & x_1 & x_2 & x_3 \\ C & \pi_1 x_1/\pi_2 & \bullet & x_4 & x_5 \\ G & \pi_1 x_2/\pi_3 & \pi_2 x_4/\pi_3 & \bullet & x_6 \\ T & \pi_1 x_3/\pi_4 & \pi_2 x_5/\pi_4 & \pi_3 x_6/\pi_4 & \bullet \end{array}$$

and x_i's are + *ve* for $i = 1, \ldots, 6$ and the diagonal elements are solutions o f eqn. (2) for given x_i and π_k.

When the process is reversible, the probability of having base $\underline{i}$ at time 0, and observing base $\underline{j}$ at time $\underline{t}$ is the same as that of base $\underline{j}$ at time 0 and observing $\underline{i}$ at time $\underline{t}$. Then,

$$f_{ij}(t) = \pi_i P_{ij}(2t) \quad \text{for } 1 \leqslant i, j \leqslant 4;\ t \geqslant 0$$

because $\pi_i Q_{ij} = \pi_j Q_{ji}$.

Thus, a general Markovian model without any assumption of substitution process becomes

$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & \bullet & x_1 & x_2 & x_3 \\ C & x_7 & \bullet & x_4 & x_5 \\ G & x_8 & x_9 & \bullet & x_6 \\ T & x_{10} & x_{11} & x_{12} & \bullet \end{array}$$

where $x_i \geqslant 0 \qquad i = 1, \ldots, 12.$

Substitution of a Sequence

Extending this model of a single site to the substitution process for an entire sequence, the model can be written $(x_l(t), y_l(t))$ denoting the bases that occur at lth position $(1 \leqslant l \leqslant n)$ where n is the length of the aligned sequence in the two species t years after divergence. Let N_{ij} be the number of times the base i in species X occupies the same position as base j in species Y. The joint distribution of N_{ij} is multinomial with n = sequence length and cell probabilities f_{ij} as in equation (1), and the substitution parameter is k. It can be shown that $k = 2\, t \sum_{i=1}^{4} \pi_i\, q_i$ since

$$\pi' Q = 0$$

The probability η that a given site contains identical nucleotides such that no substitution occured at the site, can be written as

$$\eta = \frac{\sum_i \pi_j \exp\left[-\left(q_i^x + q_i^y\right) t\right]}{\sum_i f_{ii}(t)}$$

and can be shown and tested for goodness of fit by χ^2, where

$$\chi^2 = \sum_i \sum_j \frac{(N_{ij} - n f_{ij})^2}{n f_{ij}}$$

Using the above methodology, Janzen (1985) showed that the mitochondrial DNA sequences between bovines and mice for the first and third positions were analysed with the estimate of the substitution parameter $\hat{k} = 1.089$ and variance $\hat{k} = 0.010$, for supergene 3rd base and found that the substitution rate at the third position was much higher than at the first, ($\hat{k} = 0.140$; Var $\hat{k} = 0.010$).

In the case of mice *vs* man, the probability η_i that no substitution occured was found by him for supergene as

i	*Base*	*Model* (*K*)
1	*A*	0.85
2	*C*	0.93
3	*G*	0.10
4	*T*	0.44

(Model *K* is alternative asymmetric model of Tavare)

The large differences in the substitution parameters of the bases indicate the need for more theoretical work using alternative stationary processes and MLE methods for simultaneous comparison of more than two species to construct evolutionary trees.

Randomness of Subsequences and Gene Function

It is important to assess the non-randomness of subsequences by studying the frequency distributions of these subsequences and its relation to gene function, and to search for patterns to understand the rates of evolution over time and the regulatory function of the gene (Gentleman and Mullin, 1989). Exact tests for significance of the hypothesis of randomness and the overlap capability of different lengths of subsequences are available and were used by them. The above analyses can be used to define the domain of gene function, to understand the use of short subsequences by some enzymes as signals to recognize and express nucleic acids for further genetic analysis, and even the functionability of a nucleotide sequence. The analysis of repeated subsequences may permit identification of locations of insertions, if any, into the sequence and alien DNA in particular.

By the use of Markovian processes and analysis of frequency distributions and related statistical methods mentioned above and already available, it is possible to estimate the substitution rates of a single site, a given segment of DNA sequence, the probability of the same base occupying the same position in 2 or more species, and the location of insertions, and the non-randomness of sequences for a better understanding of gene structure and function. The analysis of patterns/repeats may reveal the nature of action of endonucleases, sites of cleavage and size of fragments generated in RFLP studies. Since the substitution rates determine the mutation rates, the action of some mutagens on a set of related loci could also be studied.

Another effect of differences in substitution rates will be alteration in the proportions of repetitive and non-repetitive DNA which is important in speciation. Evidence is available on non-random (very low proportion of 0.07) of a sequence in *m*-RNA which indicates the need for further work on such material to examine the structural and functional roles of such repeats (Shukla and Srivastava, 1985).

Statistical Analysis of RFLP Data

Restriction fragment length polymorphism (RFLP) studies are now widely used in biotechnology. The procedures of these analyses are to be considered for the choice of an appropriate statistical model and therefore are summarised below. The availability of restriction enzymes is valuable in DNA manipulation. Nearly 500 endonucleases are used in RFLP analyses and involve 4—6 base pair recognition sequences. Each endonuclease is specific in its recognition sequence and conditional probabilities are to be utilised to analyse the resultant fragments. Using gel electrophoresis, size fractionation is done and individual fragments are identified by Southern blotting and hybridization to cloned radio-labelled homologous sequences. Following the standard procedures of denaturation and nick translation, individual fragments are identified by autoradiography using radio-labelled DNA as a probe. All these steps in these studies can be subjected to simple statistical modelling and analysis using proper conditional and marginal distributions.

RFLP analysis is important since restriction sites are actual samples of nucleotide sequences, the variation for the presence of sites being used to estimate the genetic divergence between individuals. Thus RFLP provides finer analysis of divergence and is superior to the use of gene frequencies data. Linkage relationships can also be determined since the DNA sequences that hybridize to a given probe constitute discrete chromosomal loci (Harvey and Muehlbauer, 1989). Thus, the fragments can be used as genetic markers. Using these markers, one can estimate the degree of variation in genetic collections, monitoring purity of hybrid seeds, selection of useful traits by exploiting the linkage relationship for a genetical analysis of quantitative characters, identification of products of cell fusion, and also in the analysis of foreign genes introduced into plant system. The methodology described earlier under substitution rates can help analyse the effect of the introduced foreign genes and their linkage with the genes of the recipient species. The level of polymorphism for a given species can be examined by using a variety of enzymes and is a useful supplement to isozyme data as is done in tomato, maize, and now in humans also (Chyi *et al.*, 1986; Neve and Belles, 1989; Raelson and Grant, 1988). An excellent example of RFLP study is the analysis of ribosomal RNA locus in tomato which is a highly repeated tandem unit. The rapid developments in RFLP technique permit the genetic mapping even to detect moderate linkages in small progenies by choosing an enzyme that gives the largest difference in fragment size between the two parents and also digests the progeny DNA. Such an analysis can be extended to recombinant DNA using the statistical procedures described in the previous section and for detecting the location of the insertion.

Polymorphism of mitochondrial DNA of higher animals can be linked to the nucleotide substitution rates, described earlier, for a better understanding of the evolution of genes and proteins. The statistical methodology of locating alien DNA insertion can be used to estimate the stability of the insertion in the new genetic background which is being attempted now in tomato by Chyi *et al.* (1986). The same statistical procedure can also be used to locate the restriction cleavage sites. The evolutionary aspects of these restriction cleavage sites and the phylogenetic relation between man and ape now being attempted can complement the diversity studies by Chakraborty (1983).

The use of endonucleases in the analysis of cytoplasmic male sterility utilising mitochondrial DNA in sorghum, tobacco and petunia and recently in pearl millet, lentil and tomato enabled detection of differences between fertile and sterile cytoplasm using the electrophoretic patterns of the fragments (Lee *et al.*, 1989). Statistical analysis of such data by multivariate techniques can help in defining the mechanisms and patterns of action of groups of endonucleases and the interaction of specific genotypes to the restriction enzymes. Studies are also reported in the estimation of linkage between restriction fragment length, isozyme pattern and morphological markers in lentil (Harvey and Muehlbauer, 1989). Similar work was also done with specific loci of tomato (Young and Tanksley 1989). However, the available method of analysis of small samples for multivariate data by non-parametric methods like Jackknife and boot strapping procedures will provide better estimation. The multivariate models for measuring overlap which is being done in ecological studies, (Lu, Smith and Good, 1989) using similarity methods can also be used in RFLP studies of endonucleases actions on possible overlaps. If several missing values are met with as in medicine, care should be exercised in modifying the methodology and interpretation (Federer and Murty, 1986: Murty and Federer, 1984).

RFLP data can supplement the estimates of divergence times between species using Markov processes described earlier. A comparison of different improved methods of MLE, LSE and method of moments for the estimation of divergence times between two or more species as proposed by Padmadisastra (1989), can be compared with the present estimates from RFLp data. The entropy measures proposed by Rao (1982) can also be extended to nucleotide substitution rates and assessment of divergence times.

The study of phylogenetic relationships between species and sub-species by RFLP analysis of chloroplast or mitochondrial DNA and a similar examination of protoplast fusion products for data available simultaneously for genetic loci as in several crops (Nevo and Belles, 1989; Raelson and Grant, 1986) can be subjected to traditional multivariate techniques as well as graphics and clustering procedures instead of the currently popular dendrograms which are empirical.

A word of caution before generalising RFLP data is the specificity of each endonuclease, hence appropriate conditional distributions and the joint distribution of the data from a set of these enzymes is to be worked out. This will also identify the common and or diversity of patterns of their action and such data need to be related to gene regulation and function. Such results can also be subjected to discriminant analysis to classify these enzymes for a set of action variables. The genotypes also can be classified for their responses using the analogy of world genetic collections.

Quadratic Discriminant Analysis and Regularization Procedures

The classical linear discriminant analysis initiated by Fisher (1936) has been widely used in germplasm classification, (Murty, 1983) and selection in crops, animals and man. The assigning of an individual to one of several groups based on *p* variables and minimising of misclassification risk particularly in small samples in high dimensional settings is of wider range of significance in several fields like medical diagnostics, space technology and economics. Rapid improvements in statistical methodology have been achieved like quadratic discrimination, regularization problems, and estimation of misclassification rates in small samples, particularly with variables of high analytical cost as in medicine and space science. The relationship between linear discrimination analysis (LDA) and quadratic discrimination analysis (QDA) is elgantly brought out by Friedman (1989) for small sample data with unequal covariance matrices, with a regularization procedure to minimize misclassification risk. Let

$$f_{\hat{k}}(X)\,\pi_{\hat{k}} = \max_{1\leqslant k<K} f_k(x)\,\pi_k$$

where $f_k(X)$ is class conditional density and π_k is the unconditional prior probability of observing a class k member. The corresponding risk function is

$$R(\hat{k} \mid X) = \frac{\sum\limits_{k=1}^{k} L(k, \hat{k}) f_k(x) \pi_k}{\sum\limits_{k=1}^{k} f_k(x) \pi_k}$$

by choosing $\hat{k}$ to minimize risk.

As $f_k(X)$ is unknown, sample observations are used to estimate $f_k(X)$. In classical LDA, the classification risk is based on

$$f_k(x) = 2\,\pi^{-1/2} \mid \Sigma_k \mid^{-1/2} \exp\left[\, - Y_2 (X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k)\right]$$

where μ_k and Σ_k are class k $(1 \leqslant k \leqslant K)$ population mean vector and cov matrix respectively. If we classify based on the following $d_k(X)$ for kth class with

$$d_k(x) = \underbrace{(X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k)}_{D^2 \text{ between } X \text{ and } \mu_k} + \ln \mid \Sigma_k \mid - 2 \ln \pi_k \qquad (2)$$

the classification rule (2) becomes QDA as the boundaries between classes are quadratic. When the quardratic terms in (2) cancel each other, it becomes LDA and $\Sigma_k = \Sigma$.

When sample sizes are small and p is large, the discriminant score is heavily weighted with the smallest eigen values biased towards low, while large ones are biased towards high. This bias becomes more pronounced as sample size decreases. Friedman (1989) proposed correction for this distortion by a regularization procedure with three alternatives, using two parameters λ and ν and by reducing the variances associated with sample estimates. The first procedure involves the replacement of individual sample covariance matrices by their averages In the second, a regularization parameter λ, $(0 \leqslant \lambda \leqslant 1)$ with $\lambda = 0$ giving QDA and $\lambda = 1$ yields LDA. In the third, called regularised discriminant analysis (RDA) further regularization is possible with two parameters λ and ν, $0 \leqslant \lambda \leqslant 1$; $0 \leqslant \nu \leqslant 1$ with λ as covariance mixing parameter and y is eigen value shrinkage parameter.

λ and ν are chosen to jointly minimize the misclassification risk

$$d_k(X) = (X - \bar{x}_k)^T \Sigma_k^{-1} (\lambda, \nu) \; (X - \bar{x}_k) + l_n \mid \hat{\Sigma}_k (\lambda, \nu) \mid - 2 \ln \pi_k \qquad (3)$$

The change due to this regularization process becomes clear when (3) is compared with (2). If ($\lambda = 0$, $\nu = 0$) we get QDA, and if ($\lambda = 1$, $\nu = 0$) we get LDA. By holding ν fixed at 0, and varying λ we can produce a range of models between QDA and LDA :

$$\hat{\sum_k}(\lambda, \nu) = (1 - \nu)\hat{\sum_k}(\lambda) + \frac{\nu}{p}\,\mathrm{tr}\left[\hat{\sum_k}(\lambda)\right] I$$

Thus, the bias in the estimates of eigen values is counteracted by choosing an optimal pair ($\hat{\lambda}$, $\hat{\nu}$) as explained above. Moreover, RDA is invariant under rotation and is scale invariant also when $\nu = 0$. With RDA, the misclassification risk is improved in small samples and unequal population cov matrices as frequently encountered in biotechnology and space science.

Elimination of variables is also possible by stepwise selection using stepwise regression program by choosing the variate which reduces the residuals as much as possible instead of random elimination as demonstrated by Weiner and Dunn (1966).

While classical discriminant analysis is mostly based on multivariate normality assumption in practice, procedures are now available for the use of discrete variables and combinations of both continuous and discrete variables using the familiar optimality arguments as in Bahadur's model of log-linear method or use of orthogonal polynomials. A new approach based on distributional distance proposed by Matusita is becoming popular (Goldstein and Dillon, 1978).

Kernel discriminant analysis which is a dynamic combination of statistics and pattern recognition and the method of nearest neighbourhood are useful non-parametric procedures in discriminant analysis (Hand, 1982). Their application will be valuable in the frontier sciences. Intensified use of Jackknife method and Boot strapping procedure to estimate misclassification rate (Effron, 1983), to reduce bias and provide approximate confidence intervals is anticipated in biotechnology and for better discrimination in small samples. Thus, a range of statistical procedures useful in several areas of science, are available.

Statistics of Size and Shape in Biology and Their Use in Space Technology

The early work of multivariate analysis in biology was oriented to interpretation in terms of size and shape (Rao, 1952) and much later in evolutionary studies (Murty, 1983), the fundamental properties of interest being the configuration of points which are elements of data set. Emphasis is now given for similar work in space science for modelling optimal size and shape of the space craft. Principal component analysis in linear systems is found to be useful to understand the components of controllability, observability and model simplification (Moore, 1981).

While applying multivariate analysis in space technology, one should note that shapes do not naturally lie in Euclidean space and imposition of multivariate normality assumption may not hold (Kendall, 1984). This difficulty was resolved by Kendall through the use of a geometry of shape manifolds, procrustean metrics, and complex projective spaces. He developed the associated distribution theory of nodes and edges natural to an analysis of shape as follows :

Let $(z_1, z_2, \ldots, z_n)$ be $\underline{n} \geqslant 3$ i.i.d. planar points in the complex plane, at most $(n - 1)$ of which are coincident. Using polar coordination, the shape of the $\Delta z_1, z_2, z_j$ can be written as (r_{j-2}, ϕ_{j-2}) for $j = 3, \ldots, n$. Then, the shape of the $\underline{n}$ points can be represented as $(r_1 \phi_1, \ldots r_{n-2} \phi_{\underline{n}-2})$. Small (1984) used the joint distribution of these shape coordinates, to define, the points (U, V, W) which lie on the sphere of radius $\frac{1}{2}$ about the origin in R^3. This theory of shape analysis can be used in biology to determine the mechanisms of shape changes during growth for a dynamic perspective of development. Treating a small change in shape as a tangent vector, multivariate analysis of changes in shape differences between groups can be done and tested using Hotelling's T^2.

Similarly, data on proportions of size and shape can be used to test differences between groups using Hotelling T^2 as shown by Campbell and Hosimann (1987). This is possible because the conditional distribution of proportions is Dirichlet with parameters that depend on size. A similar application of the above two cases to space science can be used to develop controllability measures which in turn depend on size and shape and their proportions. This can also be applied to submarines and ocean-going vessels. The data on reactions of biological material including humans, in spacecraft can also be reduced to a few supervariables.

As changes in shape and size in biology are linked to growth, the recent developments of generalized multivariate analysis of several types of growth curves with special covariance structures lend easily to prediction and estimation of future $\underline{p}$-dimensional observations using their past observations as demonstrated by Lee (1988). He has also shown that his serial-structure based predictor is superior to the simple least squares predictor and predictor based on uniform structures. Although growth curve analysis is biological, the methodology proposed by Lee could be used for predictions of technology substitutions and in space science also since in both cases future values are predicted using the corresponding past observations.

Medicine and Biometry

The advent of high powered computers has accelerated theoretical developments for utilising large data available and for advancing biomedical applications. I shall try to illustrate this with two examples of graphical structures : (a) non-parametric methods, and the other (b) by Bayesian approach. Both cases can be used for improving prediction of disease, its causal factors and progress and can permit timely remedial measures. These methods are of particular use in coronary diseases, cancer and epidemiology of fatal contagious and also plant pathogens and insect pests. Friedman and Raffsky (1982) extending the generalised correlation coefficient concept of Kendall, developed two statistics providing distribution-free tests of independence (or association) and also sensitive to general alternatives. These measures are based on interpoint distance graphs (KMST, i.e. K-minimal spanning tree well known in graph theory) for which computational methods are readily available and applicable in multivariate situations.

The concept is to obtain a statistic to measure the predictability of a random vector Y from a random vector X. Using the interpoint distance based graphs, the degree of closeness to the corresponding two vectors in one space is to be matched by closeness to the the corresponding two vectors in the other space. The above graphs have the sample observations as nodes in the Euclidean space and each node pair defines an edge. If two graphs share no edges, they are orthogonal. Thus a KMST has $K(N-1)$ edges.

In an example of epidemiology, the X space is a two dimensional map of the locations of the onset of the disease and Y space is the one-dimensional time of the onset of the disease. Thus, a high association between positions in space and time is a measure of the epidemicity. The test statistic, i.e. association measure is Γ_1 and the measure of prediction is Γ_2 the permutation distributions of both being asymptotically normal under certain conditions. Thus Γ_1 is the number of edges in the intersection of the two graphs G_x and G_y, G_x being the graph defined over the x observations and the corresponding C_y is over the y observations.

Let

$$a_{ij} = \begin{cases} 1 & \text{if edge } (i, j) \in G_x \\ 0 & \text{otherwise,} \end{cases}$$

$$b_{ij} = \begin{cases} 1 & \text{if edge } (i, j) \in G_y \\ 0 & \text{otherwise} \end{cases}$$

or define $b_{ij} = R_i(j)$

Then

$$\Gamma_1 = \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}\, b_{ij}$$

is the number of edges in the intersection

$$\Gamma_2 = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}\, b_{ij}$$

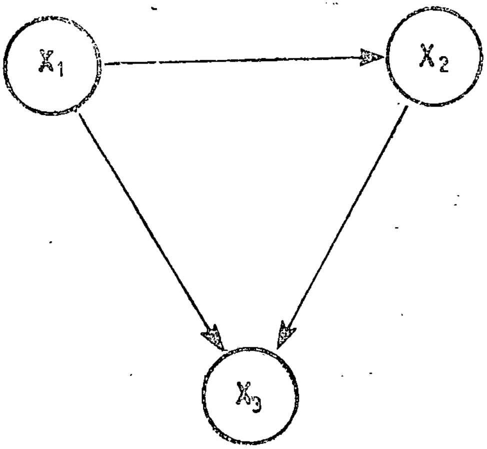
$= \Sigma_{(i,j)} \in Gx\ R_i\,(j)$ can be used for testing goodness of fit.

Thus, Γ_1 is the association measure sensitive to general alternatives and Γ_2 is used for testing goodness of fit. Using simulation studies with 100 observations/experiment, and 100 experiments/run, they demonstrated the effectiveness of this approach.

Knowing Γ_1 and Γ_2 the locations and time of onset of the disease are known to take corrective measures in outbreaks particularly those with short incubation periods, as some viruses in man and animals including poultry and pathogens like rusts and mildews in crops.

The second example is an analysis of patterns of influence using causal network and developing computationally feasible probabilistic methods. In biology and medical science in particular, we are concerned with building believable models of the phenomenon for which Bayesian belief network with the model based on substantive ideas of causality is necessary instead of assuming an empirical based linear structure. In such a framework, a coherent probabilistic assessment of causality and the multivariate distribution is explicitly built in rather than assumed.

Let $(P x_1, \ldots, x_n)$ be the causal model of a disease, some of the x_i's can be observed, and others either are not observable or may exist only as useful latent variables. For example, the data on a new patient could be incomplete, relationships between variables inexact and the terms not precisely defined. The statistician can try to exploit as much as possible the nature of the expressed variables, to "handle the uncertainty in a coherent probabilistic way". Thus, the knowledge base along with the assumptions is to be separated from the data available on a new patient, with the inference being the control mechanism of applying the knowledge to the new case. This would require the assignment of a complexity of the conditional distributions to the x random variables, defining of their joint distributions and the marginal probabilities under particular conditioning events of test results. This can be simply shown pictorially



$$P(x_1, x_2, x_3)$$

$$= P(x_2/x_1)\ P(x_1)\ P(x_3/x_1, x_2)$$

and all vertices, edges, paths, cycles, walk and directed edges defined.

Applying these concepts, Lauritzen and Spiegelhalter (1988) developed a causal model for a single muscle of a neurological disease diagrammatically described as MUNIN (Muscle and Nerve Inference Network), with the direction of causality being from left to right.

$$\left.\begin{array}{c}\text{Disease States}\\ \text{Disease node}\end{array}\right\}\ \left.\begin{array}{c}\text{Lab a \textit{priori} Distribution}\\ \text{Pathophysiological disorders}\end{array}\right\}\ \begin{array}{c}\text{Force}\\ \text{EMG}\\ \text{findings}\end{array}$$

In each box of the network, marginal probabilities are represented by histograms (under any combination of test results). Each box representing a random variable must have all conditional distributions assigned. The 'Force' variable in the above example needs 45 distributions. Based on this information, the joint distribution can be defined and the network can be solved for its marginal probabilities under particular conditioning events of the test results.. The finding of the marginal distributions is a non-polynomial time computation called NP-Hard Problem. As net work size increases, the time for finding the marginals becomes too large and simulation is the only way to tackle such a situation. With the choice of a net work with important variables, the patterns of influence among the sets of variables can be understood better. Such an analysis presented for electromyography can be used for coronory diseases, cancer and probably AIDS where our present understanding of the causal relations is not quite clear.These methods are equally useful in satellite data on weather forecasting for contingency planning as in drought-prone areas and optimal allocation of resources even in perennials as suggested by Pugliese (1988).

Information Technology-Data Reduction and Analysis

Vast atmospheric data from satellites, remote sensing data on crops, environment and land surfaces, soil survey data, All India Coordinated Schemes and several surveys (economic, demographic and health) are available with the concerned departments. However, even after the availability of advanced computers, their processing and utilisation by statisticians is not proportional to the effort in collection. Problems of quantification of measurements and data reduction also arise. Multidimensional scaling and use of other multivariate techniques and simulation studies as was used successfully in interpreting complex and large electrophoretic data (Booth and Walden, 1989) in the polypeptide synthesis in maize embryos can be applied to the above areas of remote sensing and environmental monitoring. Data availability is treated by some as information. After data is processed, interpreted, inferences made and projections given based on the analysis can only be called information. Such information is only transferable for assimilation and execution.

In the application of available methodology, several situations could arise for modifying the model and methodology (Gnanadesikan and Kettenring, 1984). Hence, an interaction in both research and teaching between theoreticians and applied users is essential and more relevant in the case of emerging fields of science as mentioned in the beginning of my talk.

In this lecture, I have only reminded a distinguished group like yours what is already known, but the impact of statistical application and related theoretical research could be much more than evident now. I wish to conclude by recalling the advice of our late President Prof. Sarvepalli Radhakrishnan that "Zeal for research and zest for teaching are inseparable." Therefore, regular training programmes and theoretial research for real world situations are essential now for the advancement of statistics and emerging fields of science. A fundamental research wing of statistics on a permanent basis devoted entirely for the above research and training activity is an immediate necessity. Inter-institutional collaboration as successfully demonstrated in the recent R.C. Bose Memorial International Symposium in New Delhi in December, 1988, should be an indispensable component of this wing. Institutions like IASRI, ISI, IISc, ISRO, CSO, TIFR and universities should be involved and invited by the Government with adequate support instead of ad-hoc measures. All of us, as scientists have a responsibility in this effort which in my opinion, will be an appropriate tribute to Dr. Panse,

REFERENCES

[1] Boothe, J. G. and Waldon, D. B. (1989) : Multivariate analysis of polypeptide syntehesis in developing maize embryos, *Theor. Appl Genet.* 77 : 495-500.

[2] Campbell, G. and Mosimann, J. E. (1987) : Multivarite methods for proportional shape, *Proc. Amer. Stat. Assn.*, Section Graphics, San Francisco August 1987, pp 10-17.

[3] Chakraborty, R. (1983) : Genetic distance and gene diversity. Some statistical considerations. *In* ; P. R. Krishnaiah (ed.), *Multivariate Analysis* VI : 77-96, North Holland, New York.

[4] Chyi, Y. S., Jargensen, R. A., Godlstein, D., Tanksley, S. D. and Figueroa, F. L., (1986) : Location and stability of Agrobacterium mediated T-DNA insertions in Lycopersicon genome, *Mol. Gen. Genetics.*, **204** : 64-69.

[5] Effron, B. (1983) : Estimating the error rate of a prediction rule : Improvement on cross validation, *JASA*, **78** : 316-331.

[6] Federer, W. T. and Murty, B. R. (1986) : Uses, limitations and requirements of multivariate analyses for inter-cropping experiments. *In* : I. B. Macneill and G. J. Umphrey (ed.), *Proc. Symp. in Statistics and Festchrift in honour of V. M. Joshi*, D. Reidel Publishing Co. Boston, U.S.A., pp. 269-283.

[7] Fisher, R. A. (1936) : The use of multiple measurements in taxonomic problems, *Ann. Eugenics*, **7** : 179-188.

[8] Friedman, J. H. (1989) : Regularized discriminant analysis, *JASA*, **84** : 165-174.

[9] Friedman, J. H. and Rafsky, L. C. (1982) : Graph-theoretic measures of multivariate association and prediction, *Ann. Statist.* **11** : 377-391.

[10] Gentleman, J. F., and Mullin, R. C. (1989) : The distribution of the frequency occurrence of nucleotide subsequences on their overlap capability, *Biometrics* **45** : 35-52.

[11] Gnandadesikan, R., and Kettenring, J. R. (1984) : *In* H. A. David and H. T. David. (ed), *Statistics : An appraisal*. Proc. 50th Anniv. Conf. Iowa State Statistical Laboratory, Iowa State University Press, pp. 309-337.

[12] Goldstein, M. and Dillon, W. R. (1978) : *Discrete Discriminant Analysis*, Edn. 1, John Wiley Sons, pp. 1-186.

[13] Hand, D. J. (1982) : *Kernel Discriminant Studies*, Research Studies Press, Letchworth, Herts, England, pp. 1-252.

[14] Harvey, M. J., and Muehlbauer, F. J. (1989) : Linkages between restriction fragment length, isozyme and morphological markers in lentil, *Theoret. Appl. Genet.*, **77** : 395-400.

[15] Janzen, T. (1985) : Estimation of substitution rates for Homologous DNA sequences, *CSU Tech. Rept.* **163**.

[16] Kendall, D. G. (1984) : Shape manifolds, procrustean metrics and complex projective spaces, *Bull. London Math. Soc.*, **16** : 81-121.

[17] Lauritzen, S. L. and Spiegelhalter, D. J. (1988) : Local computations with probabiliiies on graphical structures and their applications to expert systems, *J. R Statist. Soc.* (*B*) **50** : 157-224.

[18] Lee, J. C. (1988) : Prediction and estimation of growth curves with special covariance structures, *JASA*, **83** : 432-440.

[19] Lee, S. H., Muthikrishnan, S., Sorenesen, E. L. and Liang, G. L. (1989) : Restriction endonuclease analysis of mitochondrial DNA from sorghum with fertile and sterile cytoplasms, *Theoret. Appl. Genet.*, 77 : 379-382.

[20] Lu, R., Smith, E. P. and Good, I. J. (1989) : Multivariate measure of similarity and niche overlap, *Theoret. Pop. Biol.* 35 : 1-21.

[21] Mellor, J. W. (1980) : Agriculture in growth-Changing research and data needs for effective policy, "Dr. V. G. Panse Memorial Lecture". Indian Society of Agricultural Statistics, New Delhi, Feb. 1980-pp. 1-16.

[22] Moore, B. C. (1981) : Principal component analysis in linear systems, controllability, observability and model reduction, *IEET Trans. Autom. Control*, AC-26 : 17-32.

[23] Murty, B. R. (1983) : Interdisciplinary research in genetics, Presidential Address, *Ind. J. Genet.*, 43 : 113-122.

[24] Murty, B. R. and Federer W. T. (1984) : Missing observations in multivariate analysis, Technical Series, Bu-860M Biometrics Unit, Cornell University, Ithaca, New York, pp. 1-21.

[25] Nevo, E. L. and Belles, A. (1989) : Genetic diversity in wild emmer wheats in Israel and Turkey, *Theor. Appl. Gen.*, 77 : 421-447.

[26] Padmadisastra, S. (1989) : Estimating divergence times, *Theoret. Pop. Biol.*, 34 : 297-319.

[27] Pugliese, A. (1988) : Optimal resource allocation in perennial plants. A continuous time model, *Theoret. Pop. Biol.*, 34 : 215-247.

[28] Raelson, J. V. and Grant, W. F. (1988) : Isoenzyme data used for evaluation of origin of Lotus, corniculatus, *Theoret. Appl. Genet*; 76 : 267-276.,

[29] Rao, C. R. (1952) : *Advanced Statistical Methods in Biometrical Research*, Edn. 1, John Wiley and Sons, London.

[30] Rao, C. R. (1982) : Diversity and dissimilarity coefficients. A unified approach, *Theoret. Pop. Biol.*, 21 : 24-43.

31] Shukla, R. and Srivastava, R. C. (1985) : The statistical analysis of direct repeats in nucleic acid sequences, *J. Appl. Prob.*, 22 : 15-24.

[32] Small, C. G. (1984) : A classification theorem on planar distributions based on shape statistics of independent tetrads, *Math. Proc. Comb. Phil. Soc.*, 96 : 543-547.

[33] Tavare, G. (1986) : The general reversible model of the distribution of base frequencies, *JASA*, 80 : 405-417.

[34] Weiner, J. M. and Dunn, O. J. (1966) : Elimination of variates in linear discrimination problems, *Biometrics*, 22 : 268-279.

[35] Young, N. D. and Tanksley, S. D. (1987) : Restriction fragment analysis and the concept of graphical genotypes, *Theoret. Appl. Genet.* 77 : 95-101.

[36] Young, N. D. and Tanksley S. D. (1989) : RFLP analysis of the size of chromosomal segments retained around Tm-2 locus of tomato during back cross breeding, *Theoret. Appl. Genet.*, 77 : 353-359.