Statistical Directions in the 1990s

Joseph Gani - Santa Barbara, California, USA

1. Introduction

A volume of appreciation for Sir Ronald Fisher (Fienberg and Hinkley, 1980) was recently re-issued on the occasion of the hundredth anniversary of his birth in 1890. The book reviews his contributions to what may now be considered the classical elements of mathematical statistics, namely analysis of variance, foundations of theoretical statistics, randomization and design, statistical estimation, analysis of categorical data and discriminant analysis. Fisher died in 1962, but his influence remained substantial well into the 1980s, as did also Neyman's (b.1894, d.1981) and Wald's (b.1902, d.1950).

It is interesting to note that in 1990, *Mathematical Reviews* still uses the following classifications for Statistics: foundations, multivariate analysis, sufficiency, regression and correlation, decision theory, experimental design, sampling theory, sample surveys, sequential methods, distribution theory, inference from stochastic processes, parametric inference, non-parametric inference, engineering statistics, applications. There has clearly been much activity in these areas over the past 20 years; but excepting for inference from stochastic processes and engineering statistics, one might well ask if much had changed on the statistical scene since the Fisher-Neyman era.

In probability, which is regarded as the basis of statistical theory, a standard text of the 1960s was the book by Loève (1960). This consisted of six main parts on: elementary probability theory, notions of measure theory, general concepts and tools of probability theory, independence, dependence, and elements of random analysis.

The 1990 classifications of *Mathematical Reviews* refer not only to Probability but also to Stochastic Processes, already indicating a change from the narrower purview of Loève's text. These sections include for Probability: foundations, probability theory on algebraic and topological structures, combinatorial probability, stochastic geometry, random sets, distribution theory, limit theorems, and for Stochastic Processes: stochastic analysis, Markov processes, special processes.

There has been a distinct shift, over the past several years, towards the probabilistic analysis of random processes varying in time, as well as in a random environment. Inference from such processes has presented statisticians with a variety of novel and sometimes difficult problems.

I propose to discuss two important developments which have taken place during the past 20 years. Both have occurred within the context of greatly increased computing power, and the ready availability of powerful PCs. They are robust statistical methods, and the use of Markov processes in probabilistic modelling. After giving simple examples of these, I shall suggest some possible directions of interest to statisticians in the 1990s.

2. Robust statistical methods

A departure from the principles of Fisher and Neyman which gathered momentum during the 1970s was the concept of robustness. The idea was simplicity itself: it consisted of developing statistical methods which were insensitive to changes in the basic model. Statistical estimation, for example, is based on a set of measurements $X_1, X_2, \ldots, X_n$, where the underlying distribution is often assumed to be normal $N(\mu,\sigma^2)$. The sample mean $\bar{X}$ is the maximum likelihood, best unbiased, minimax and asymptotically efficient estimate of μ. But any small departure from the basic assumption of normality will lead to poor performance of the estimate of location: this is an undesirable property.

Tukey (1977), Huber (1981) and Hampel et al. (1986) endeavoured during the 1970s and 1980s to develop precise robust procedures for estimation, inspection sampling plans, regression, smoothing, change-point models and experimental design in statistics. We illustrate such robust statistical methods in the very simplest case.

It is already assumed that measurements X_i, $i = 1,\ldots,n$, are $N(\mu,\sigma^2)$. We may think of each measurement as $X_i = \mu + e_i$, where $e_i \sim N(0, \sigma^2)$; $\bar{X}$ would then be an optimal estimator of μ. Suppose that measurements are now subject to a malfunction with probability ε, independently of any error occurring in the absence of a malfunction. Then the distribution function (d.f.) of the e_i would be

$$G(x) = (1-\varepsilon)\,\Phi\left(\frac{x}{\sigma}\right) + \varepsilon H(x),$$

where Φ is the standard normal d.f., and H(x) can be any other d.f.

The tails of G are likely to be heavier than those of the normal d.f., and some of the e_i would tend to be outliers. Thus, one could expect that $\bar{X}$ might lead to inaccurate estimates of μ, since $\bar{X}$ could be biased if G were not symmetric, and the variance of $\bar{X}$ might be much higher than when no malfunctions occur.

The lack of robustness of $\bar{X}$ as an estimate of μ is well known, and is usually dealt with by rejecting outliers. But it is also recognised that the median and the trimmed mean will provide location estimates less affected by outliers. There are several approaches to the problem of the robust estimation of the mean; one of them is Huber's asymptotic minmax. If G is assumed symmetric about 0, and σ^2 is known, Huber has proposed minmax estimates which are generalisations of the maximum likelihood estimates. Huber's estimates also turn out to satisfy Hampel's influence functions criteria.

Any difficulties encountered in calculating these estimates, or equivalent quantities in other robust procedures, are greatly reduced by computerised techniques for which programs are already available. Calculations which deterred statisticians in the 1960s can now be carried out with relative ease, and these have encouraged the use of computer intensive methods not only in implementing robust statistical procedures, but in all other areas of statistical methodology.

3. Markov processes in modelling

A lively development in probability and its applications since the publication of Loève's text has been the analysis of various Markov processes and their use in modelling natural phenomena. For a random rainfall process X_t varying in time, one may fit a model having a somewhat loose correlation structure; such is the Markov process

$$X_{t+1} = \rho X_t + \varepsilon_t, \quad t = 0,1,2,\ldots,$$

where $-1 \le \rho \le 1$, and $\varepsilon_t \sim N(0,\sigma^2)$, say. Such a process will give a reasonable approximation to the time series of recorded daily rainfall in a city, from which ρ and σ can be estimated. The structure of the process may not provide a causal explanation of the phenomenon, but it is sufficiently precise to allow some predictions to be made about future rainfall.

For other random processes, such as spatial epidemics, a greater degree of structure may need to be built into the model. Consider trees planted in an orchard on a lattice, where an infection may be transmitted from a diseased tree I to its nearest susceptible neighbour S (see Figure 1), where each lattice point has 8 nearest neighbours. Some diseased trees I may also recover from the infection and become immunes R. We may assume that during time (t, t+1), the following transition probabilities will obtain for any susceptible S, or infective I:

$$P\{S \to I \mid k \text{ infective neighbours}\} = k\beta, \; k = 0,1,\ldots,8; \; P\{I \to R\} = \gamma.$$

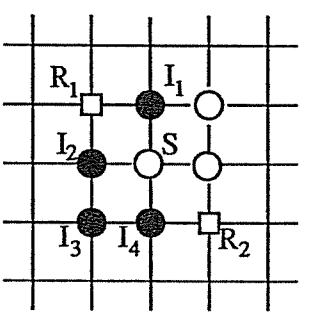

FIGURE 1
Transmitting infection from I_i (i = 1,2,3,4) to S

We are now faced with a Markov field (cf. Adler, 1981) for which it is rather difficult to derive too many analytic results, as Durrett (1988) has indicated. However, we may use computer methods to obtain a wide range of simulated answers to questions of infectious spread.

For example, it can be shown that a concentrated block of n infective trees is likely to infect fewer susceptibles than a randomly spread set of n infectives distributed over the lattice. Further, by ringing a set of infectives by a large enough proportion of immunes, one can effectively control the spread of infection. Here, computer intensive methods guide one's intuition, and allow one to draw both practical and theoretical conclusions which were simply not possible before the current generation of powerful computers and PCs.

4. Directions for the 1990s

Can we make some informed guesses about likely developments in statistics and probability during the next decade? Whatever the research thrusts may turn out to be, we can be fairly certain that they will rely heavily on computers and PCs, whose power and versatility will continue to increase enormously in the 1990s. PCs in particular have now become flexible and efficient instruments, small enough to be portable, and cheap enough to be affordable by any serious researcher. This means that heavy statistical analysis and complex calculations can now be carried out with relative ease, often using available programs; algorithmic methods and simulation have come into regular use, and frequently provide answers to problems, or at least directions in which to search. Empirical distributions may be used instead of the classical normal or the negative exponential, for example. There is already a large literature on the "bootstrap", a method developed by Efron (1982), using observed data to generate further simulated data, as often required in adaptive procedures. Here, the data itself determines the model, rather than being forced to fit a particular theoretical structure; for further details see Hall (1990).

If one were to suggest an area of continuing growth in statistics, one would probably select inference on stochastic processes, including time series and Markov fields. The principles of inference on time-dependent processes X(t) in continuous time are complex, the more so as one rarely knows for certain whether a process is stationary or not. Problems in multiple time series $\underset{\sim}{X}(t)$ and in Markov fields will prove even more challenging: the effects of non-stationarity, the variation due to a random environment and the estimation of change points, will pose sets of important problems. We have already encountered such a process with a change point threshold in the "Greenhouse Effect"; how can one be certain that such an effect exists, and determine at what point in time global warming has begun? The amount of data to be analysed is enormous, and the techniques highly sophisticated.

A possible area of growth in probability could well be the analysis of Markov fields, and the development of limit theorems applicable to them. Much progress has already been made on Ising type models in physics, and on percolation processes (see Kesten, 1982) in recent years. The field remains a very active one, not only in its theoretical, but also its applied aspects, such as image processing, spatial epidemic spread, and genetic trees.

While several of the directions outlined may seem somewhat theoretical, they are in fact all applicable, and in some cases very practical. The 1990s will be a decade in which government, commerce and industry will place increasing emphasis on the collection, analysis and evaluation of large sets of data - social, economic and scientific. Economic and environmental surveys are already commonplace and likely to become more so. It is not too difficult to foresee the social pressures which will dictate the general orientation of statisticians, nor to estimate the influence of computers and PCs on how data will be analysed and evaluated.

What appears much more difficult to predict is where the statistical manpower is likely to come from. It seems fairly likely that by the year 2000, there will be a chronic shortage of well-trained statisticians and probabilistic modellers, not only in colleges and universities, but also in every walk of government, commerce and industry. It is already clear that we are not training an adequate number of statisticians to replace those of us currently at work; if we do not respond to this serious challenge, and plan a massive statistical training programme, we may fail the most crucial of all our tests in the 1990s.

References

Adler, R J (1981) *The Geometry of Random Fields.* John Wiley, Chichester.

Durrett, R (1988) *Lecture Notes on Particle Systems and Percolation.* Wadsworth and Brooks/Cole, Monterey, CA.

Efron, B (1982) *The Jacknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia, PA.

Fienberg, S E and Hinkley, D V (eds) (1980) R A Fisher : An Appreciation. *Lecture Notes in Statistics*, No 1. Springer-Verlag, New York.

Hall, P G (1990) *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York. (To appear)

Hampel, F R, Ronchetti, E M, Rousseeuw, P J and Stahel, W A (1986) *Robust Statistics : The Approach Based on Influence Functions.* John Wiley, New York.

Huber, P J (1981) *Robust Statistics.* John Wiley, New York.

Kesten, H (1982) *Percolation Theory for Mathematicians.* Birkhauser, Boston, MA.

Loève, M (1960) *Probability Theory.* Van Nostrand, Princeton, N.J.

Tukey, J W (1977) *Exploratory Data Analysis.* Addison-Wesley, Reading, MA.