INDIAN STATISTICAL INSTITUTE

TWENTYSECOND CONVOCATION ADDRESS

# The Fascination of Statistics

**Sir David Cox, F.R.S.**

*Imperial College of Science and Technology, London*


**14th January 1988**

203 BARRACKPORE TRUNK ROAD
CALCUTTA 700 035

# INDIAN STATISTICAL INSTITUTE

TWENTYSECOND CONVOCATION ADDRESS

# The Fascination of Statistics

*by*

**Sir David Cox, F.R.S.**

*Imperial College of Science and Technology, London*

INDIAN STATISTICAL INSTITUTE LIBRARY
No. 028899
Date 10 FEB 1988
CALCUTTA

**14th January 1988**

203 BARRACKPORE TRUNK ROAD
CALCUTTA 700 035

Mr. President, Mr. Chairman, Professor Rao, Director, Graduating Students, Guests and Colleagues :

It is my first task to congratulate very warmly all graduating students on the completion of their work. You are graduating from an Institution with a fine international reputation and I join in wishing you successful careers, rewarding to you personally and to Society.

My second task is to express my great pleasure and appreciation at being here in Calcutta, both at being present at this Ceremony today and also at having the great honour of giving Lectures in memory of Professor P. C. Mahalanobis, that commanding figure on the Indian and international scientific scene.

Your Director has told me that a technical address is usual at Convocation. The opportunity of expressing enthusiasm for one's subject to a general audience must always be welcome. I think virtually all scientists feel their own field to be very specially challenging and important and, of course, statisticians are no exception. It is hard to express the fascination of a scientific subject without entering into fairly intricate detail, and in this respect one is at some disadvantage as compared with academic colleagues studying such fields as literature or music, where the matters at issue can, sometimes at least, be explained fairly readily to a general audience.

I would like to try and explain the special interest and fascination of statistics as a field of study, partly by expounding some current and continuing themes and partly by explaining some of the differences from many other areas of work.

The first and obvious point to make about the subject is the enormous breadth of application. The issues concerned with the collection and interpretation of data about the economic and social aspects of life, very much in the minds of the pioneers of our field, remain of enormous importance, but I shall in these remarks concentrate on the role of statistics in science and technology. In contrast to the extreme specialization of so much modern science, summed up in the phrase "getting to know more and more about

less and less", statistical ideas arise in a great many contexts. Thus my own work in recent months has brought me into contact with

> medical workers concerned with the (so-called) quality of life experienced under various treatment regimes ;
>
> hydrologists and meteorologists concerned with short-term patterns of rainfall ;
>
> psychologists concerned with the behaviour of individuals confined in institutions ;
>
> epidemiologists concerned with the spread of infectious diseases ;
>
> fishery research workers concerned with patterns of development of the whale population in the Antarctic ;
>
> technologists interested in the relation between process variability and industrial quality and reliability;

and this list is by no means complete.

Now all these investigations are very different, not just in the obvious sense of being concerned with disparate areas of activity, but also in the kind and quantity and quality of data it is possible to obtain and the difficulties of its collection and interpretation. What then are the common features which mean that there is some intellectual unity to the study of these issues, different though the techniques of analysis are in detail ?

Two key notions, which in a sense have a certain duality, are those of variability and uncertainty. The positive aspects of the first are, of course, celebrated in the motto of your Institute. Variability is a phenomenon in the physical world to be measured, analysed and where appropriate explained. By contrast, uncertainty is an aspect of knowledge, incidentally not at all to be equated with ignorance. The great Polish-English novelist Joseph Conrad wrote at some point words to the effect that in the end all is ignorance ; I suspect he meant uncertainty not ignorance. Now the theory of probability provides a mathematical framework for describing uncontrolled variability and, in a different role, a basis for measuring uncertainty. The

philosophical interest and importance of the subject stem from this dual claim to be able to study and analyze random variability and also to be able to come to terms with uncertainty, to recognize its existence, to measure it and to show that advancement of knowledge and vigorous action in face of uncertainty are possible and rational. Recognition of uncertainty does not imply nihilism ; nor need it force us into what the Americans sometimes call one-handedness.

Now I do not want to give the impression that statisticians spend most of their time pondering on these big philosophical issues ; much more time is spent wondering whether such and such a graph means what it seems to, or whether all reasonably practicable precautions have been taken to avoid biases in data collection. Such issues are sometimes not highly technical, more a matter of experience and commonsense, but anyone who dismisses them as uninteresting or trivial is making a great mistake.

So far I have talked about issues that have been around a long time. That they are still a focus of much discussion and research is a comment both on their ubiquity and on their difficulty. Let me comment now on some of the issues that are currently the focus of much research. Incidentally something approaching 100,000 pages of new statistical research is published each year, so that any attempt to select a few topics can represent only a very personal choice. I will choose just three items for brief comment.

First there is the matter of basing statistical analysis on models (i.e., idealized mathematical representations) that are specific to the scientific issue under study ; this is to be contrasted with the extreme empiricism of much statistical work in which representations are used which are of very wide applicability and which therefore almost inevitably are not strongly linked to the detailed "science" of the problem.

As an example, consider the analysis of rainfall, taken on a short time scale, say one-hour rainfalls or even on a shorter time interval still. Large amounts of data are involved in any study of this kind ; one wants many years of hourly rainfalls, preferably at many sites over an area, say a river basin, and this is a lot of numbers. One of the most celebrated sets of data of this kind contains about $14 \times 10^6$ observations. One could summarise

such data in a descriptive fashion in various ways, but for some deeper understanding some kind of mathematical model is useful. Such a model can be of at least three different kinds. At one extreme are the models of dynamic meteorology, large systems of nonlinear partial differential equations aiming to represent with some realism the physical mechanisms involved. These require a very large computer for their solution in real time and, as I understand it, give good results, at least under the climatic conditions of Western Europe, for periods of a few days. At the other extreme are totally empirical statistical models, i.e. ones with no explicit physical interpretation. These are essentially generalizations of familiar regression models. Now my coworkers, Professor Rodriguez-Iturbe from Caracas in Venezuela and Dr Isham from University College London and I have been investigating models that are in a sense intermediate. We consider a stochastic point process of rain cells of random durations and depths ; in a spatial temporal model these cells move randomly in the plane before "dying". These models involve a fairly small number of adjustable parameters, usually 5 or 6, and so are vastly simpler than the models of dynamic meteorology, but the parameters have a physical interpretation and the models do, among other purposes, serve as a basis for reducing large amounts of data in a meaningful way. This application is in some ways typical of the development of special models. The study of the models is via a mixture of classical mathematical analysis and computer simulation.

The above remarks illustrate another very important focus of current statistical work, namely the development of methods for handling very large quantities of data. It is commonplace that the computer has made many kinds of analysis of data much easier ; in some ways it is even more significant that the collection and storage of vast quantities of data has become possible. In the physical sciences, notably in the earth sciences, enormous international data-bases are being assembled. In a single clinical trial there may be 5,000 patients interviewed quarterly for four or five years and on each visit a substantial number of biochemical, physical and other information may be recorded. In many fields much of the data collected with so much care is never seriously analysed and it is a major issue for the future to develop sensitive and practicable methods for handling such large amounts of data.

Contact with applications is crucial for these developments and it is worth stressing that many of the more mathematical developments in the subject also stem from specific applications or groups of application. As with other areas of science in which mathematics is used, the types of mathematical argument that can be usefully deployed in a statistical context get ever more varied and advanced. Currently there is, for instance, much interest in the application of differential geometry in a statistical setting, something in which Professor C. R. Rao is a pioneer. While I believe that statistics has to be seen as primarily an applied subject, there is much need for powerful mathematicians in the field.

This is particularly relevant to the rather technical area of higher order asymptotic theory, an area in which your Director is one of the world's leading experts and which is the principal subject of my lectures here. The study of asymptotic expansions has a distinguished history in pure mathematics, the rigorous theory being especially associated with the name of H. Poincaré. In classical applied mathematics such methods are widely used, in particular in connection with analytically intractable problems in fluid mechanics. There is a long history in probability theory too, but it is only within perhaps the last 10 years or so that these matters have become quite pressing in statistics at a relatively applied level, stemming essentially from the computer making possible much wider use of relatively complex methods and the handling of large quantities of data, in the way I sketched above. This has led to a substantial body of advanced mathematical work, some rethinking of concepts and the hope for the development of quite simple elegant results of broad applicability.

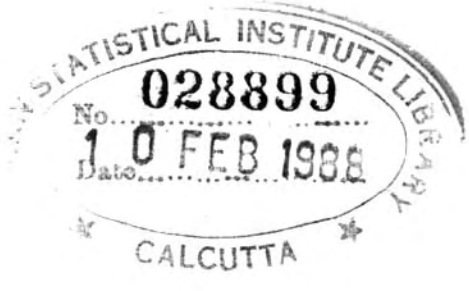
It is of special significance, I believe, that the use of the computer, while it has drastically changed the strategy of dealing with particular problems, has not at all diminished the need for *mathematical* skills. Another important development which I shall mention only very briefly concerns the design of computerized 'expert systems' for statistical analysis. I see these as interesting primarily because they force theoretical statisticians to think not only about the details of particular techniques but also about the strategy for their effective implementation ; this applies as much to the important areas of design of experiments and surveys as to their analysis.

There is not time to say very much about the difficult educational problems raised by the teaching of statistics. I want to mention just one important issue. This is that it is likely that many nonspecialists in many different fields of work, by no means confined to science and technology, will at some point have to look at arguments or reports with some statistical content. It is therefore very desirable that a great many people should have some background so as to be able to understand statistical work in broad outline ; it is not a question of mastering detailed statistical techniques but rather of understanding some key ideas. This raises educational issues quite different from those involved in training statistical specialists, where in many ways the critical issue is that of striking some suitable balance between theory and application.

I hope that these few remarks have, in particular, given those of you who are not specializing in statistical work some feeling for the breadth and interest of the subject.

I repeat my very warm good wishes to all graduating students.

Thank you for listening to me.


STATISTICAL INSTITUTE LIBRARY
No. 028899
Date 10 FEB 1988
CALCUTTA