

## The Nature of Statistical Inference Johann Bernoulli Lecture 1997

D.R. Cox

*Department of Statistics and Nuffield College, Oxford*

### 1. INTRODUCTION

Bernoulli is one of the great family names in the mathematical sciences and it is in particular totally appropriate that the major international society for Probability and Mathematical Statistics is called the Bernoulli Society. It has been very interesting to hear of the connection with Groningen and I greatly appreciate the invitation to give this lecture and to visit the University again.

I want to talk, so far as possible in a nontechnical way, about the nature of theories of statistical inference. My viewpoint will be eclectic, a word whose dictionary meaning is

*selecting beliefs etc. from various sources; attached to no particular school of philosophy.*

This viewpoint is, I believe, the proper one for our subject and is, in particular, very much in line with the attitude of the great Dutch mathematician and statistician D.van Dantzig, who, in the 1950's wrote two influential papers attacking what he strikingly called statistical priesthood. Nevertheless it rules out an approach which has much appeal to those with a mathematical outlook, namely the formulation of a single set of axioms from which all else follows.

Because the audience for this lecture is not primarily statistical let me first say a word about the broad field now sometimes called statistical science. Its mathematical foundation is largely, although not quite entirely, the theory of probability, regarded as a chapter in pure mathematics, and statisticians concern themselves with broadly three interlocking areas

- the application of probability models, for example to genetics, to epidemics, to operational research (queueing, etc), and to a number of topics in the physical sciences, and so on
- some general principles for the design of investigations, including the design of experiments, the design of observational studies and of sampling methods
- some general principles for the analysis and interpretation of empirical data.

The formal theories of inference which are the subject of this lecture are part, but only a part, of the third of these. The discussion is largely motivated by the natural and social sciences and by science-based technology, rather than by issues of public affairs which gave the subject of statistics its in some ways rather misleading name.

## 2. TWO BROAD THEMES

The subject is partly about the subtle interplay between two different but intimately related themes

- unexplained variability
- uncertainty.

Unexplained variability is an almost universal feature of observations of the real world. Think of the weather, of survival times of patients with a particular disease, of the educational performance of children, and of almost any kind of measurement in the laboratory sciences. If we want to represent that unexplained variability mathematically, of course in idealized form, one main approach is to turn to the theory of probability building on a notion of stability of long-run frequencies. This is a phenomenological use of probability.

Uncertainty is about our knowledge. Will it rain in Groningen tomorrow? We do not know for sure. Can we attach a probability to the event that it will rain? This is very closely connected to, but not quite the same as, the issue of modelling the variability of rainfall. Most statistical inference is concerned with a more ambitious problem involving a more subtle form of uncertainty: broadly, what is the uncertainty involved in drawing conclusions from a particular investigation currently under analysis?

For example, a clinical trial comparing the survival times of patients randomized, subject to informed consent, between a new treatment  $T$  and the best current therapy  $C$ , appears to show some advantage to  $T$ . Is that firmly established? Within what limits is the gain in survival likely to lie? Let us be clear that many other issues, especially of design, arise and would be of concern to a research worker in this field, but the above will be enough for this lecture!

Can we use probability to assess uncertainty in such contexts? Here we want to use probability epistemologically.

## 3. A LITTLE HISTORY OF PROBABILITY

Until the middle of the 19th century probability, and indeed statistics, were part of the main stream of the mathematical sciences, as shown by the contributions of, for example, the Bernoullis, Laplace and Gauss. In the early years of the 20th century, although distinguished work continued, probability theory fell in some disrepute in the mathematical world, essentially because of doubts about the interpretation of probability. Although not the first attempt to axiomatize the subject, by far the most influential was that of A.N. Kolmogorov in 1933. This has been the basis of most mathematical work since, a remarkable flowering of probability theory into a beautiful topic within modern pure mathematics. Issues of interpretation do not arise.

When, however, we turn to applications to statistical inference the question of the meaning of probability cannot be escaped!

## 4. FORMULATION

We commonly address the theory of statistical analysis in the following framework.

We have data  $y_1, \dots, y_n$  representing what we shall call the *responses* of interest; for example these might be survival times of  $n$  individuals in a clinical trial. In addition there will be vector explanatory variables  $x_1, \dots, x_n$  representing such features as treatment used, medical history of the individual, etc.

Most formal discussion of statistical inference proceeds from the following highly nontrivial abstractions:

- the responses are regarded as values of random variables  $Y_1, \dots, Y_n$
- we have a family of possible probability distributions for these random variables, in general depending on the  $x$ 's, which we call a family of *models* for the problem. These represent the phenomenon of observed variability
- questions of interest are formulated as questions about these distributions, i.e. the data are regarded as of interest only in so far as they tell us about the underlying probability distribution which is intended to describe the population or random system that generated the data.

In particular, in so-called parametric formulations the family of models is of known mathematical form and is determined by one or more unknown constants, called unknown *parameters*. The objective of statistical analysis is then formulated as

- to draw conclusions about the unknown parameters
- to assess whether the family of models is adequate or needs revision or indeed to be abandoned.

There are other possible objectives, notably the making of decisions or predictions, but there is not time to discuss these.

##### 5. AN EXAMPLE

Here is a not entirely fictitious example for illustration.

Suppose that the survival times of 10 patients suffering from a particular severe disease are

0.4, 1.6, 0.1, 0.2, 3.0, 1.6, 0.9, 0.3, 3.9, 0.4 years.

Suppose also that on the basis partly of experience of similar problems and partly of simple stochastic models, it is assumed that the survival times are independently distributed with an exponential probability density,  $\mu^{-1}e^{-y/\mu}$ . Here  $\mu$  is interpreted as the mean survival time that would be obtained from a very large population of individuals under the same conditions.

What conclusions can we draw about the parameter  $\mu$ , the population mean? The mean of the data is 1.26 years, but it would be absurd to conclude that  $\mu$  is exactly the same. How close is it likely to be? If in the research literature on this subject a large body of data had given a mean of 0.8 years,

are our data reasonably consistent with that? More generally if we are comparing two groups  $T$  and  $C$ , what uncertainty is involved in drawing a conclusion about the difference between them; what steps can be taken in design to reduce that uncertainty?

Of course real examples are nearly always much more complicated.

For at least the last 200 years the way to tackle such issues has been at a fundamental level a matter of dispute. That has not prevented the development of an enormous body of widely accepted methods of analysis and design which are an integral part of the techniques of research in many fields.

To an appreciable extent the discussion hinges on the meaning of probability in such an inferential context, i.e. on its role to describe uncertainty.

## 6. THE MAIN APPROACHES

It is a helpful simplification to think of four broad approaches which can be described by the kinds of mathematical problem that arise and by the conceptual base employed. These are as follows:

- the rather ill-defined but rich set of ideas that is conveniently called Fisherian, after the geneticist and statistician R.A. Fisher. Note that Fisher explicitly emphasized the need for a variety of approaches for different problems; he was dismissive of axiomatic arguments
- an approach strongly based on a frequency theory of probability, initially developed to explain Fisher's ideas more concretely, and emphasizing operational concepts. This, the Neyman-Pearson theory, is that most commonly presented in text books and courses
- an approach in which probability represents rational degree of belief, i.e.  $P(A \mid \text{data}, H)$  measures the degree of belief that a reasonable person would have in  $A$  given the data under analysis and additional information represented by  $H$ . This goes back to Laplace and before; in its modern form it is associated with the philosopher R. Carnap and especially with the geophysicist and applied mathematician H. Jeffreys
- an approach in which probability represents *your* degree of belief constrained by a requirement of selfconsistency. This is associated with F.P. Ramsey, I.J. Good, B. de Finetti and L.J. Savage. There is no assumption that different people faced by the same evidence have the same probability.

For convenience I shall name the first three approaches after respectively Fisher, Neyman and Pearson and after Jeffreys and call the last personalistic probability. Van Dantzig's criticisms of priesthood were directed primarily at the first and last of these, in fact at Fisher and Savage.

The adjective Bayesian ought to refer to any use of a particular simple and entirely uncontroversial result in elementary probability theory; in fact it nowadays tends to refer either to the last two or sometimes even just to the last of these approaches to statistical inference.

As I have already implicitly indicated there is merit in all these approaches in different contexts.

## 7. A QUICK SKETCH

It is hazardous to describe and compare four subtly different approaches in a few words but let me try!

In the first two approaches we regard  $\mu$  as an unknown constant capturing in idealized form properties of the system under study. We develop methods of analysis that gain their justification by their performance, i.e. frequency properties, under hypothetical repetition. This is in line with the general scientific principle that measuring devices are to be calibrated by considering their performance when used. I assume general familiarity with notions of confidence limits and significance tests, so-called  $p$  values, interpreted along the lines of their hypothetical frequency properties in repeated sampling.

The differences between the Fisherian and the Neyman-Pearson approaches in some senses may seem quite minor. They are essentially as follows:

1. in the Fisherian approaches emphasis is placed on what he called the simple test of significance, on the likelihood function and on such principles as sufficiency, and, most controversially, on interval estimation through fiducial distributions
2. the Neyman-Pearson approach emphasizes operational requirements such as power and other explicit indicators of sensitivity
3. the Neyman-Pearson approach tends to use a somewhat decision oriented terminology, such as accepting and rejecting hypotheses
4. there is a difference in the attitude to what philosophers call the problem of the unique case

In our specific problem the likelihood, defined as the probability of the observations considered as a function of the unknown parameter, is

$$\mu^{-1}e^{-v_1/\mu} \dots \mu^{-1}e^{-v_n/\mu} = \mu^{-n}e^{-\Sigma v_j/\mu},$$

thus depending on the data only via the sum or equivalently the mean. In the Neyman-Pearson theory this dependence on the mean is deduced from some optimality requirement such as maximum power.

The next step in both approaches is the mathematical one of finding the probability distribution of the random variable  $\bar{Y}$  representing the sample mean. In fact  $\bar{Y}/\mu$  has a special well tabulated distribution corresponding to one tenth of a random variable having the chi-squared distribution with  $2 \times 10$  degrees of freedom.

It follows that for example

$$P(\bar{Y}/\mu > 1.42) = 0.1,$$

$$P(\mu < 0.70\bar{Y}) = 0.1.$$

On replacing the random variable  $\bar{Y}$  by its observed value of 1.26, we obtain what in Neyman-Pearson theory is called a lower 90 percent confidence interval.

The difference of approach that is most subtle and which, sometimes at least, has practical importance is the problem of the unique case. Consider for illustration the question : what is the probability that it will rain in Groningen

tomorrow? Suppose that the proportion of a very large number of May days with rain is  $p$ . Now strictly speaking in Neyman-Pearson terms, probability refers only to repetitive phenomena. We may say it will rain tomorrow. We are then following what Neyman called an inductive rule of behaviour such that in repeated applications we will be right a proportion  $p$  of times. This is regarded as less than saying the probability of the unique event is  $p$ . In Fisher's view it is allowable to say that the probability is  $p$  that it will rain tomorrow provided that

- there is a frequency interpretation
- there is appropriate conditioning.

That is we must not be able to recognize tomorrow as belonging to some subensemble with a different frequency, i.e. we do not have or decide not to use further information. Similarly, when we use probability to assess the uncertainty in drawing conclusions from a particular set of data, the probabilities must be relevant to the conditions of that investigation. The difficulties of formalizing that idea mathematically are one primary source of conflict between Fisherian and Neyman-Pearson analyses. (In many instances the conflict does not arise.)

From Fisher's standpoint it is entirely right, in the absence of further evidence, to make the above probability statement about  $\mu$ ; he was, however, clearly wrong in supposing that such individual probability statements could be manipulated by the ordinary laws of probability theory.

A rather contrived example of the conflict over conditioning is this. Suppose that in designing the above study I was unsure whether to take 10 individuals or 20 and that I tossed a coin to decide which. In the ensemble of repetitions of the investigation half the time there would be 10 and half 20. Is this relevant to interpreting the data which I actually have, where I know there were 10. In Fisherian theory we condition on the 10 but this is certainly not automatically so in Neyman-Pearson theory. There are other more subtle and realistic applications of the same idea.

A delicate issue in theoretical formulations is how to formulate a conditioning principle. A delicate issue in practical cases is to determine how far to condition; overconditioning can, in various senses, lead to an effective loss of information.

Jeffreys's approach to the issue of inference has exactly the same target as Fisher's; what can we reasonably learn about a parameter, say  $\mu$ , in the light of a model for the data. Jeffreys argues, however, that a different notion of probability, that of reasonable degree of belief, is needed to achieve this. Arguments are produced as to why this measure obeys the same mathematical laws of probability theory, a by no means obvious conclusion. Then if  $m$  is any particular value of  $\mu$  a simple formula, often called Bayes's rule, gives that

$$P(\mu = m \mid \text{data}) \propto P(\text{data} \mid m) \times P(\mu = m).$$

Note that strictly the probabilities for  $\mu$  are probability densities.

The terms in this relation are called respectively the posterior distribution of  $\mu$ , the likelihood and the prior distribution of  $\mu$ . The last is concerned with knowledge of  $\mu$  in the absence of the data. In Jeffreys's approach to estimation, essentially following Laplace, this is taken to be very dispersed, representing lack of knowledge, and Jeffreys produces invariance arguments for taking it to be proportional to  $m^{-1}$ . This leads to numerically the same answer as the confidence interval approach.

Two difficulties with this are

- how do we deal with prior information that is not statistical in the narrow sense?
- is the notion of lack of knowledge represented by a very flat distribution, preferably uniquely defined for each problem, a viable one?

Almost all current workers in this field consider the answer to this last question to be, *no*. It appears to be the case that the Jeffreys- type priors in a low number of dimensions lead to conclusions broadly compatible with Fisherian and Neyman-Pearson analyses but that in a large number of dimensions manifestly unacceptable results may arise.

For that reason the great majority of current work using Bayes's theorem as a basis for statistical inference pays at least formal lip service to yet another notion of probability, that of personalistic degree of belief, measuring how strongly *you* believe in, say, *A* given information *H*. This is in principle elicited from *your* behaviour in hypothetical betting games depending on *A*. It can be shown that *your* probabilities can be defined in such a way that any departure from the laws of probability theory is equivalent to a so-called incoherent assignment, in which in effect inferior betting strategies would be used.

The idea is then that *you* should elicit *your* prior probabilities and then apply Bayes's theorem to find *your* posterior distribution.

Note the following points

- no distinction is drawn between a probability that is based on carefully studied evidence and one that is just a guess
- the approach can be helpful for private assessment but to be helpful for public enlightenment it has to be supplemented either by arguments justifying the prior or by studies of robustness
- the theory provides an elegant qualitative representation of the adaptation to new evidence
- it does not, however, cover simply the possibility of a conflict between prior information and the data or the closely associated possibility that a wholly new formulation needs to be found
- the arguments about coherency apply at one time and do not necessarily demand that the prior is constant in time and in some very special circumstances the prior may quite reasonably depend on the data
- if the prior is based directly on empirical data the whole process is, at least in principle, uncontroversial

- clearly judgement is involved in overall assessments of information and the approach formalizes this. It does not, at least directly say how to form probabilities based on information that is not specifically statistical; *you* may elicit what *your* probability currently is but are given no direct guidance on what that probability would sensibly be.

On this last point, the theory does not tell *you* how to assess *your* probability that exposure to residential EMF fields induces childhood leukemia given the information that cellular and tissue level experiments and animal experiments suggest no effect and that the combined analysis of 10-20 international epidemiological studies has an ambiguous interpretation. It is arguable that, assuming it is at all useful to quantify such notions, it is desirable to find, of course inevitably very roughly, an empirical frequency of correctness in somewhat similar situations; see Fisher's definition of probability mentioned above.

While, as noted above, most current theoretical Bayesian thinking rejects the notion of representing ignorance, much current Bayesian applied work uses flat priors at some point.

#### 8. SOME KEY QUESTIONS

In considering the relative advantages of these different approaches the following are some central issues:

1. in many practical contexts the numerical conclusions from different approaches will be very similar, once the primary formulation is agreed
2. all approaches other than the personalistic Bayesian one deal essentially with uncertainty arising via statistical variability. Clearly there are other kinds of uncertainty
3. are all probabilities comparable? This is a fundamental tenet of an axiomatic approach to statistical inference based on personalistic probability
4. there is an element of personal judgement in the interpretation of scientific and technological data. Is this best separated from relatively objective aspects? How useful for public discussion is an approach focussing on personal behaviour?

#### 9. CASE FOR ECLECTICISM

Why is an eclectic view valuable?

- some frequency notion of probability is needed to represent variability in the real world
- the calibration of measures of uncertainty via their (hypothetical) frequency performance in application is in line with general principles of measurement
- the Fisherian notions of likelihood, sufficiency and especially conditioning have direct appeal from many viewpoints
- the Neyman-Pearson operational viewpoint gives support to Fisherian principles, is the simplest basis for discussing issues such as robustness and non and semiparametric methods and points the way also to decision-based formulations

- the notion of reasoned degree of belief seems a sound formulation of objectives in many contexts although there are major problems in quantitative application especially where information that is not statistical in the narrow sense is involved
- the notion of personalistic probability recognizes the element of judgement involved in interpretation, provides a formalism for its incorporation, although it gives no clear guide on how to proceed in converting nonstatistical information into probabilities and does not distinguish probabilities as to whether they are or are not based on solid evidence.

#### 10. CONCLUSION

There are many aspects not covered in this short survey. I hope I have given you some insight into the issues involved. As I have stressed, the numerical answers from the different approaches tend to be similar, at least once a formulation of a model is agreed. The importance of the issues lies more in the impact on the kinds of question we are led to ask and on our perception of the purposes of analysis and interpretation. Also there are many critical aspects to the statistical analysis and interpretation of data that are outside the formal theory of inference.

I have deliberately concentrated on the conceptual issues involved rather than on their mathematical formulation and development. Had I aimed to show the mathematical challenge and interest of statistics I would, of course, have given a quite different lecture! The principles I have been discussing have to be formulated mathematically in sufficient generality that the specific problems of an enormously and fascinatingly wide span of applications can be addressed.

I am grateful to the Leverhulme Foundation for an Emeritus Fellowship and to the University of Padua for its hospitality during the writing of some of this lecture. It is a pleasure also to thank Professor W. Schaafsma for much encouragement and advice.

## ADDENDUM. SOME CRYPTIC ISSUES FOR DISCUSSION

1. How is overconditioning to be avoided?
2. How convincing is A. Birnbaum's argument that likelihood functions from different experiments that happen to be proportional should be treated identically (the so-called strong likelihood principle)?
3. What is the role of probability in formulating models for systems, such as economic time series, where even hypothetical repetition is hard to envisage?
4. Should nonparametric and semiparametric formulations be forced into a likelihood-based framework?
5. Is it fruitful to treat inference and decision analysis somewhat separately?
6. How possible and fruitful is it to treat quantitatively uncertainty not derived from statistical variability?
7. Are all sensible probabilities ultimately frequency-based?
8. Was R.A. Fisher right to deride axiomatic formulations (in statistics)?
9. How can the randomization theory of experimental design and survey sampling best be accommodated within broader statistical theory?
10. Is the formulation of personalistic probability by de Finetti and Savage the wrong way round? It puts betting behaviour first and belief to be derived from that behaviour last.
11. How useful is a personalistic theory as a base for public discussion?
12. In a Bayesian formulation should priors constructed retrospectively after seeing the data be treated distinctively?
13. Is the only sound justification of much current Bayesian work using rather flat priors the generation of (approximate) confidence limits? Or do the various forms of reference priors have some other viable justification?
14. What is the role in theory and in practice of upper and lower probabilities?