Distributional inference

Albers, Casper

Rijksuniversiteit Groningen

Distributional Inference: The Limits of Reason

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
vrijdag 7 maart 2003
om 14.15 uur
door

Casper Johannes Albers

geboren op 30 juni 1975
te Hengelo (Overijssel)

Promotor:
Prof.dr. W. Schaafsma

Beoordelingscommissie:
Prof.dr. H.G. Dehling
Prof.dr. R.C. Jansen
Prof.dr. A.W. van der Vaart

Preface

Stripped to its essentials, statistics is about inferences and decisions on the basis of numerical evidence which is statistical in the sense that *things could have been different*. Which inferences and decisions are 'most appropriate', given some initial ideas and a set of data? That is the question to which the mathematical statistician restricts his attention. The applied statistician has a more comprehensive task. He will have to participate in the discussion about the design of the experiments, the choice of the data to be evaluated, the interpretation of the inferences, etcetera.

It is true, and completely natural, that the 'procedures' proposed by mathematical statisticians are based on principles which can be modified. In practice, the relevant data should be sufficiently abundant to accept the corresponding inferences or decisions as 'reasonable'. If factual information is scarce or irrelevant or not trustworthy then one should not rely on the precise results prescribed by such procedures. Anybody should be aware of what Kant called *the limits of reason* ('die Grenzen der Vernunft').

Procedures for making inferences on the basis of data are usually based on a mathematical model which comprises a specification of the probabilistic context within which the data arise, and a specification of the inferential or decision-making context.

Probability statements and distributional inferences are of particular interest, e.g. as an intermediary between data and decision. 'The making of statistical inferences in distributional form is conceptionally complicated because the epistemic 'probabilities' assigned are mixtures of fact and fiction. In this respect they are essentially different from 'physical' or 'frequency-theoretic' probabilities. The distributional form is so attractive and useful, however, that it should be pursued'[1].

In Part I of this thesis it will be made very clear, by elaborating on examples, that the precise probability statements and distributional inferences prescribed by some 'rational' mathematical-statistical procedure are sometimes not relevant. If, however, the factual information is sufficiently abundant, then the inferences based on it deserve to play a part in the discussion. 'As workers in Science we aim, in fact, at methods of inference which shall be equally convincing to all freely reasoning minds, entirely independently of any intentions that might be furthered by utilizing the knowledge inferred'[2]. Mathematical statisticians try to be as 'objective' as possible, even if they declare themselves to be subjectivists.

[1] A.H. Kroese et al., Distributional Inference, *Statistica Neerlandica*, **49**:1, 63–82, 1995

[2] R.A. Fisher, *Statistical Methods and Scientific Inference*, third edition, Macmillan, New York, 1973, p. 107

the statistical model. Secondly, a similar example, the two-envelopes problem, is considered. Again, the difficulties involving the numerical specification of conditional probabilities are in the forefront.

The second and most important part deals with the situation where one has a random sample $x_1, \ldots, x_n$ from a distribution with density f. The goal is to use the sample to form an estimate of f or, almost equivalently, to generate a distributional inference about $y(= x_{n+1})$. A new method is discussed to estimate the density f, where 'initial knowledge' of f is incorporated in the model. This is done by specifying a probability density ψ as the 'initial guess' for f. Also the degree of confidence in this ψ is quantified and incorporated in the method. By means of a multi-modal approach, incorporating aspects from both Classical and Bayesian statistics, and on basis of the sample x, 'initial guess' ψ (and the degree of confidence in ψ), an estimate $\hat{f}$ of f is generated. When the initial guess ψ is not unreasonable, this density estimate performs better, in general, than the generally used kernel methods. This is no surprise, since the kernel method makes no use of ψ. It is at this point unclear how the comparison will turn out when ψ is incorporated in the kernel method.

To study the applicability of the developed method, an extensive data set about the pollution of Dutch waters is considered. Previous investigations showed that the different concentrations of pollutants can reasonably well be described by lognormal distributions. A complication is that the concentrations can only be measured when they are above a certain detection threshold. The density estimation theory of this thesis, adapted to mentioned complication, is used to 'fine-tune' the 'initial guess' of lognormality to the data. The resulting density estimates are better than the density estimates obtained previously by fitting lognormal densities.

The density estimation theory of this thesis can usefully be applied to the goodness of fit context where a statement is required about the truth or falsity of the hypothesis H_0: $f = \psi$. The resulting goodness of fit tests have interesting relations with the well-known χ^2-test, Kolmogorovs test, and Neymans 'smooth tests'.

To emphasize the usefulness of distributional inference, an example from the interface of multivariate analysis and time-series analysis is discussed.

In Part II two types of (distributional) inference will be examined. The first one is that of a density (e.g. the density of a future observation) which has to be estimated on basis of a combination of a sample and a priori 'knowledge'. The second one is that of a statement which has to be made about the truth or falsity of the hypothesis that the density is exactly the one specified. In the latter case an accept-reject statement may be more appropriate than the assignment of a probability.

In Part III applications are made of the theory developed in Part II, but also with a concrete problem in mind. This two-way traffic is essential. Statistics needs applications and many applications need statistics. 'As regards mathematics, you cannot separate it from its applications to the external world, and you cannot separate statistics from mathematics, or mathematical statistics from applied statistics.'[3]

The reader might wonder whether something 'new' can be found in this thesis: many of the arguments to be used were already availably half a century ago. It is the combination of such arguments which is pursued. A specific feature of Chapter 3 is that it goes one step into the direction of a Bayesian approach by claiming that an a priori guess is available. The second step, the specification of an entire a priori distribution, is not made because it does not seem appropriate. The density estimates provided seem to be 'very good', though, slight modifications can still lead to further improvements.

Multi-modal 'compromises' can, of course, differ. The rational man likes logical validity and unicity of solution. Unfortunately, such niceties are not attainable in statistics. One will study a variety of approaches to the same problem, each one resulting in an expert opinion. If these opinions are sufficiently alike then any one of them will do, a summary can be presented. If the opinions are too much different, then one should *not* present such concluding summary.

Directions for the reader: this thesis contains a report of some of the problems I encountered for the past four years. Although these problems were of interest for my own development, not all of them are of direct interest to the 'average reader'. Such cases are mentioned where they occur, and the sections concerned could be skipped from reading.

Casper Albers
Groningen, December 2002

[3] M. KAC ET AL., *Discrete thoughts — Essays on Mathematics, Science, and Philosophy*, BIRKHÄUSER, BOSTON, 1986

Acknowledgements
I am indebted to my promotor Willem Schaafsma, who was always available with his support and enthusiasm. The theory of Chapter 2 was developed jointly with Barteld Kooi. The estimator of Chapter 3 was studied before by Rob de Bruin and Diemer Salomé. Other issues were brought under our attention by or were discussed with (in order of appearance in this thesis):

Sandy Zabell (Northwestern University of Illinois) commented on the Kullback-Leibler number (Section 1.3), Rieks op den Akker (University of Twente) made some valuable remarks on the prisoner's dilemma (Section 1.9), Lammert ter Veld (Groningen) referred us to the Protagoras paradox (Section 2.2), Roland Auer and Wim Oudshoorn (Groningen) suggested a decision function in Section 2.4, Wilbert Kallenberg (Twente) and Subha Chakraborti (University of Alabama) commented on Chapter 4, Richard Davis (University of Colorado) and Thomas Mikosch (University of Copenhagen) brought the unit root problem (Chapter 6) under our attention, Jan Hulscher (Groningen) and Gerlof de Roos advised on Chapter 7, Alwin Stegeman (Groningen) gave helpful advice on printing theses, and, finally, Willem Albers (Twente) advised for the Summary and Samenvatting.

The coworkers at the *Centre for Quantitative Methods* (CQM) in Eindhoven, especially Mynt Zijlstra, Jaap Praagman, Marijke Swaving, Bert de Vries, and Ruud van Lieshout, assisted me in constructing the first drafts of Chapter 5.

R.N.M. Duin of the *Rijksinstituut voor Kust en Zee* kindly granted us permission to use the data set *Monitoring Watertoestand des Lands* (Chapter 5). G.Th. de Roos granted permission to use his ornithological data (Chapter 7).

To my former Probability & Statistics colleagues in Groningen: Alwin, Bojan, Daniel, Diemer, Evgeni, Herold, Hui, Mook, Thomas, Willem, Wout: thank you for the nice working atmosphere in the periods you were in Groningen. Finally, thanks to Marijn.

Contents