

University of Groningen

Distributional inference

Albers, Casper

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2003

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Albers, C. J. (2003). Distributional inference: the limits of reason Groningen: s.n.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

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 pursued¹.

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.

¹A.H. KROESE ET AL., Distributional Inference, *Statistica Neerlandica*, **49**:1, 63–82, 1995

²R.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, \dots, 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 available 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

³M. 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

Preface	iii
I The limits of reason	1
1 How to assign probabilities if you must	3
1.1 Predictive distributional inference, an example	3
1.2 Background information	4
1.3 A Fisher-Neyman-Pearson-Wald approach	6
1.4 A detailed discussion of some specific procedures	10
1.5 What if Player I uses a randomized strategy	12
1.6 Adapting the theory to alternative loss functions	14
1.7 Extension	20
1.8 An classroom example with Bernoulli trials	22
1.9 Related well-known examples	26
2 Trying to resolve the two-envelope problem	29
2.1 The two-envelope problem	29
2.2 Explorations	30
2.3 Probability theory	33
2.4 Mathematical statistics	35
2.5 Game theory	38
2.6 Discussion: the limits of reason	43
II Statistical inference, distributional inference in particular	45
3 Estimating a density by adapting an initial guess	47
3.1 Introduction	48
3.2 The original density estimation method	50

3.3	Asymptotic properties of the quantile function estimate	53
3.4	Asymptotic properties of the density function	57
3.5	Comparison with other methods	60
3.6	Improvements through U -statistic symmetrization	62
3.7	Simulation studies	64
3.8	Specifying the initial guess	68
3.9	Extending the theory to the bivariate case	72
3.10	Discussion	73
4	A goodness of fit test, smoother than smooth	75
4.1	Goodness of fit testing	75
4.2	Specifications	78
4.3	The extreme case $m = 1$	80
4.4	The case $m = 2$	82
4.5	The general case $m \geq 2$	85
4.6	Relation with Neyman's smooth tests	88
4.7	Relations with other goodness of fit tests	92
III	Applications	93
5	Analyzing water-quality data	95
5.1	Description of the data	95
5.2	All measurements are above the threshold	96
5.3	Some measurements below threshold	98
5.4	Substances involving multiple thresholds	101
5.5	Complications	101
6	The interface between time series and multivariate analysis	103
6.1	Introduction to the unit root problem	103
6.2	The case $n = 2$	106
6.3	Exact inferences in the case $n > 2$	110
6.4	Focusing on exact tests for the null hypothesis	111
6.5	Survey of the literature for the case $n = \infty$	113
6.6	Discussion	114

7 Making statistical inferences about a frequency unseen	115
7.1 Introduction	115
7.2 Primary approaches	118
7.3 Obtaining a solution by ignoring day effects	119
7.4 Obtaining a solution by taking day effects into account	121
7.5 Discussion	125
Appendices	127
A Proof of Theorem 4.1	129
B Critical values for the $\ f_n^{(m)} - \psi\ _1$-test statistic	133
Bibliography	137
Author index	145
Summary	147
Samenvatting	149

