BAYESIAN STATISTICS

Proceedings of the First International Meeting held in Valencia (Spain)

May 28 to June 2, 1979

Edited by

BERNARDO, J.M. (Universidad de Valencia)
DEGROOT, M.H. (Carnegie-Mellon University)
LINDLEY, D.V. (University College London)
SMITH, A.F.M. (University of Nottingham)

1980

University Press
Valencia, Spain

I.S.B.N.: 84-370-0172-2
Depósito Legal: V-883-1981
Valencia, Spain: University Press

CONTENTS

PAGS

PREFACE

1. Foundations of Subjective Probability and Decision Making

GIRON, F.J. and RIOS, S. (Universidad de Málaga, Universidad de Madrid)
Quasi-Bayesian Behaviour: a more realistic approach to decision making? 17

HILL, B.M. (University of Michigan and University of Utah)
On finite additivity, non-conglomerability, and statistical paradoxes 39

DISCUSSION

2. Sequential learning, discontinuities and changes

MAKOV, U.E. (Chelsea College, London)
Approximations of unsupervised Bayes learning procedures 69

SMITH, A.F.M. (University of Nottingham)
Change-Point problems: approaches and applications 83

HARRISON, P.J. and SMITH, J.Q. (University of Warwick)
Discontinuity, decision and conflict 99

DISCUSSION

3. Likelihood, sufficiency and ancillarity

AKAIKE, H. (The Institute of Mathematical Statistics, Tokyo)
Likelihood and the Bayes procedure 143

DAWID, A.P. (The City University, London)
A Bayesian look at nuisance parameters 167

DISCUSSION

4. Approximations

MOUCHART, M. (Université Catholique de Louvain) and SIMAR, L. (Facultés Universitaires Saint-Louis, Buxelles)
Least squares approximation in Bayesian analysis 207

LINDLEY, D.V. (University College London)
Approximate Bayesian methods 223

DISCUSSION

5. Regression and time series

BROWN, P.J. (Imperial College, London)
Aspects of multivariate regression 249

DEMPSTER, A.P. (Harvard University)
Bayesian inference in applied statistics 266

DISCUSSION

6. Bayesian and non-Bayesian conditional inference

BARNARD, G.A. (University of Waterloo)
Pivotal inference and the Bayesian controversy 295

DISCUSSION

7. Personal and inter-personal ethics

SAVAGE, I.R. (Yale University)
On not being rational 321

KADANE, J.B. (Carnegie- Mellon University) and SEDRANSK, N. (State University of New York at Albany)
Toward a more ethical clinical trial 329

DISCUSSION

8. Sensitivity to models
FREEMAN, P.R. (University of Leicester)
On the number of outliers in data from a linear model 349
BOX, G.E.P. (University of Wisconsin)
Sampling inference, Bayes inference, and robustness in the advancement of learning 366
DISCUSSION

9. Improving judgements using feedback
DE GROOT, M.H. (Carnegie-Mellon University)
Improving predictive distributions 385
PRESS, S.J. (University of California, Riverside)
Bayesian inference in group judgement formulation and decision making using qualitative controlled feedback 397
DISCUSSION

10. Predictive sample reuse
GEISSER, S. (University of Minnesota)
Predictive sample reuse techniques for censored data 433
DISCUSSION

11. Beliefs about beliefs
DICKEY, J.M. (University College of Wales, Aberystwyth)
Beliefs about beliefs, a theory of stochastic assessments of subjective probabilities 471
GOOD, I.J. (Virginia Politechnic and State University)
Some history of the hierarchical Bayesian methodology 489
DISCUSSION

12. Bayesian non-parametric theory
DALAL, S.R. (Rutgers University)
Non-parametric Bayes Decision Theory 523
DISCUSSION

13. Coherence of models and utilities
LEONARD, T. (University of Warwick)
The roles of inductive modelling and coherence in Bayesian statistics..... 537
NOVICK, M.R., DEKEYREL, D.F. and CHUANG, D.T. (The University of Iowa)
Local and regional coherence utility assessment procedures 557
DISCUSSION

14. Hypothesis Testing
ZELLNER, A. and SIOW, A. (University of Chicago)
Posterior odds ratio for selected regression hypotheses 585
BERNARDO, J.M. (Universidad de Valencia)
A Bayesian analysis of classical hypothesis testing 605
DISCUSSION

AUTHOR INDEX

CONFERENCE PARTICIPANTS

A) INVITED PAPERS

AKAIKE, H.
The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku
Tokyo, JAPAN

BERNARDO, J.M.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10, SPAIN

BOX, G.E.P.
Departament Statistics
Univ. of Wisconsin
Madison, WI 53706, U.S.A.

BROWN, P.J.
Department of Mathematics
Imperial College
Queen's Gate, London SW7 2BZ
ENGLAND

DAWID, A.P.
Department of Mathematics
The City University
Northampton Square
London EC1V OHB
ENGLAND

DEGROOT, M.H.
Department of Statistics
Carnegie-Mellon Univ.
Pittsburg, PA. 15213
U.S.A.

DEMPSTER, A.
Department of Statistics
Harvard University
1 Oxford Street
Cambridge, MA 02138
U.S.A.

DICKEY, J.M.
Department of Mathematics and Statistics
State University of N.Y. at Albany
1400 Washington Ave.,
Albany, N.Y. 12222
U.S.A.

FREEMAN, P.R.
Department of Mathematics
The University
Leicester, LE1 7RH
ENGLAND

GEISSER, S.
School of Statistics
270 Vincent Hall
206 Church Street S.E.
University of Minnesota
Minneapolis, Minnesota 55455
U.S.A.

GIRON, F.J.
Departamento de Estadística
Facultad de Ciencias
Universidad de Málaga
Camino de la Misericordia s/n
Málaga, SPAIN

GOOD, J.
Department of Statistics
Hutcheson Hall
Virginia Polytechnic Institute
Blacksburg, VA 24061, U.S.A.

HARRISON, J.
Statistics Department
University of Warwick
Coventry CV4 7AL
ENGLAND

HILL, B.M.
The University of Michigan
Department of Statistics
1447 Mason Hall,
Ann Arbor, Michigan 48109
U.S.A.

KADANE. J.B.
Carnegie-Mellon University
Department of Statistics
Schenley Park
Pittsburgh, Pennsilvania 15213
U.S.A.

LEONARD, T.
University of Wisconsin-Madison
Department of Statistics
1210 West Dayton Street
Madison, Wisconsin 53706
U.S.A.

LINDLEY, D.V.
2 Periton Lane
Minehead
Somerset, TA24 8AQ
ENGLAND

MAKOV, U.
Department of Mathematics
Chelsea College
London SW3 6LX
ENGLAND

MOUCHART, M.
CORE, University Catholique de Louvain
34 Voie du Roman Pays
1348 Louvain-La-Neuve
BELGIUM

NOVICK, M.
356 Lindquist Center
The University of Iowa
Iowa City, Iowa 52242
U.S.A.

PRESS, S.J.
Department of Statistics
University of California
Riverside, CA 92521
U.S.A.

RIOS, S.
Departamento de Investigación
Operativa y Estadística
Facultad de Matemáticas
Universidad Complutense
Ciudad Universitaria s/n
Madrid-3, SPAIN

SAVAGE, I.R.
Department of Statistics
Yale University
Box 2179 Yale Sta.,
New Haven, CT 06520
U.S.A.

SMITH, A.F.M.
Department of Mathematics
University Park
Nottingham, NG7 2RD
ENGLAND

ZELLNER, A.
H.G.B. Alexander Research Foundation
Graduate School of Business
University of Chicago
Chicago, Illinois 60637
U.S.A.

B) INVITED DISCUSSANTS

BARNARD, G.A.
Mill House, Hurst Green
Brightlingsea, Colchester
Essex CO7 0EH
ENGLAND

DALAL, S.R.
Bell Laboratories
600 Mountain Avenue
Murray Hill, New Jersey 07974
U.S.A.

DU MOUCHEL, W.H.
Department of Mathematics
Mass. Inst. Technology
Rm. 2-382
Cambridge, Mass, 02139
U.S.A.

DUNSMORE, I.R.
Department of Probability & Statistics
The University of Sheffield
Sheffield S3 7RH
ENGLAND

EDDY, W.
Department of Statistics
Carnegie-Mellon University
Pittsburg, PA. 15213
U.S.A.

FIENBERG, S.E.
Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA. 15213
U.S.A.

GOLDSTEIN, M.
Department of Mathematics
The University
Cottingham Rd.,
Hull HU6 7RX
ENGLAND

GREN, J.
Econometric Institute
Central School of Planning & Statistics
Rakowiecka St. 24
02-521 Warsaw
POLAND

GUTTMAN, I.
Department of Statistics
Univ. of Toronto
Toronto, Ontario
CANADA

JAYNES, E.T.
Department of Physics
Washington University
St. Louis, Mo. 63130
U.S.A.

O'HAGAN, A.
Department of Statistics
University of Warwick
Coventry CV4 7AL
ENGLAND

PEÑA, D.
Escuela de Organización Industrial
Ministerio de Industria y Energia
Avda. de la Moncloa, s/n
Madrid-3, SPAIN

PICCINATO, L.
Istituto di Calcolo delle Probabilita
Università di Roma
Citta Universitaria
00100 Roma, ITALY

SKENE, A.M.
Department of Mathematics
University Park
Nottingham, NG7 2RD
ENGLAND

SPIEGEHALTER, D.,
Department of Mathematics
University Park
Nottingham, NG7 2RD
ENGLAND

VILLEGAS, C.
Department of Mathematics
Simon Fraser University
Burnaby, B.C.
CANADA V5A156

C) OTHER PARTICIPANTS

ARMERO, C.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10, SPAIN

BACCHELLI, B.
Istituto di Matematica
Via Cicognara, 7
20100 Milano
ITALY

BASULTO, J.
Departamento de Estadística
Facultad de Ciencias Economicas
y Empresariales
Sevilla, SPAIN

BAYARRI, M.J.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10, SPAIN

BERMUDEZ, J.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10, SPAIN

BERTINO, S.F.A.
Istituto di Calcolo delle Probabilita
Facolta di Scienze Statistiche
Citta Universitaria
00100 Roma, ITALY

BOUGAS, I.
Bell Canada/Management Sciences
Rm. 1120
620 Belmont
Montreal, Quebec
CANADA

BRADSHAW, S.
Statistics Department
University of Warwick
Coventry, CV4 7AL
ENGLAND

BURN, R.
Department of Mathematics
Brighton Polytechnic
Moulsecoomb Brighton BN2 4GJ
ENGLAND

CASTILLO, E.
Escuela de Ingenieros de Caminos
Universidad de Santander
Avda. de los Castros, s/n
Santander, SPAIN

DI NATALE, M.
Istituto di Matematica
Via Cicognara, 7
Milano, ITALY

FERRANDIZ, J.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10, SPAIN

FRASER, D.A.S.
Department of Mathematical Statistics
Univ. of Toronto
Toronto, Ontario
CANADA M5S1A1

FRENCH, S.
Department of Decision Theory
University of Manchester, M13 9PL
ENGLAND

GAMBETTA, G.
Via Pratello, 50
Bologna, ITALY

GARCIA-HERNANDEZ, F.
821 Plum Street
Riverside, California 92507
U.S.A.

GHÈRARDINI, P.
Istituto per la Applicazioni del Calcolo
Consiglio Nazionales delle Ricerche
Via del Policlinico, 137
00161 Roma, ITALY

GILLE, P.
95, Rue Wariclet
B-5872 Corroy-Le-Grand
BELGIUM

GOMEZ VILLEGAS, M.A.
Universidad Complutense
Facultad de Ciencias Matemáticas
Ciudad Universitaria
Madrid-3, SPAIN

HOULE, A.
2400 Chemin Ste-Foy
Quebec G1V 1T2
CANADA

LAYACHI, I.
Faculté des Sciences de Rabat
Rabat, MOROCCO

JINKINSON, R.A.
20 Upper Tooting Park
London SW17
ENGLAND

JONES, S.G.
Statistics Department
Warwick University
Coventry CV4 7AL
ENGLAND

LEFORT, G.
Institut National Agronomique
16 Rue Claude Bernard
75231 Paris, Cedex 05
FRANCE

LUCEÑO, A.
Escuela de Ingenieros de Caminos
Universidad de Santander
Avda. de los Castros, s/n
Santander, SPAIN

MASCIOLI, F.
Istituto Matematico "Guido Castelmoro"
Universitá di Roma
Citta Universitaria
00100 Roma, ITALY

MCLAREN, A.D.
Department of Statistics
University Gardens
Glasgow G12 8QW
Scotland, UNITED KINGDOM

MORA, E.
Escuela de Ingenieros de Caminos
Universidad de Santander
Avda. de los Castros, s/n
Santander, SPAIN

ORSI, R.
Istituto Statistico-Matematico
Universita di Modena
Via Giardini, 454
41100 Modena
ITALY

PALLESEN, L.
IMSOR, Bygn. 349
The Technical University of Denmark
DK-2800 Lyngby
DENMARK

PAPPAS, P.D.
15, Aristoxenoo St.
Athens, 501
GREECE

PORCAR, M.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10 SPAIN

PRAT-BARTES, A.
Cátedra de Estadística E.T.S.I.I.B.
Universidad Politécnica de Barcelona
Avda. Generalísimo, 647
Barcelona-28, SPAIN

PUIG-PEY, J.
Escuela de Ingenieros de Caminos
Universidad de Santander
Avda. de los Castros, s/n
Santander, SPAIN

RABENA, M.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10, SPAIN

RAMALHOTO, M.F.
Pr. Prof. Santos Andrea, 14 - 6°
1500 Lisboa
PORTUGAL

RICCIARDI, N.
Istituto di Calcolo delle Probabilita
Universita di Roma
Citta Universitaria
00100 Roma, ITALY

RIOS, M.J.
Departamento de Estadística
Facultad de Matemáticas
Ciudad Universitaria, s/n
Madrid,3, SPAIN

ROMERO, R.
Departamento de Estadística
Escuela Técnica Superior
de Ingenieros Agrónomos
Avda. Blasco Ibáñez, 21
Valencia-10, SPAIN

SANJUAN, L.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10 SPAIN

SCOZZAFAVA, R.
Dipartamento di Matematica
Via della Montagnola, 30
60100 Ancona, ITALY

SENDRA, M.
Departamento de Bioestadística
Facultad de Medicina
Avda. Blasco Ibáñez, 17
Valencia-10 SPAIN

SLATER, M.
20 Upper Tooting Park
London SW17
ENGLAND

SOLOMON, H.
U.S. Office of Naval Research
223 Old Marylebone Road
London NW1 5TH
ENGLAND

SOUZA, R.C.
Statistics Department
University of Warwick
Coventry CV4 7AL
ENGLAND

SPEZZAFERRI, F.
Istituto di Calcolo delle Probabilita
Facolta di Scienze Statistiche
Citta Universitaria
Roma, ITALY

STEGMAN, C.
University of Pittsburgh
Pittsburgh, Pa 15260
U.S.A.

STORNELLI, D.
Via Aurelia 747
00165 Roma
ITALY

STROUD, T.W.F.
Department of Mathematics & Statistics
Queen's University
Kingston, Ontario
CANADA

SUDGEN, R.A.
Department of Mathematics
University of Nottingham
Nottingham, NG7 2RD
ENGLAND

VERDINELLI, I.
Instituto di Calcolo delle Probabilita
Universita di Roma
Citta Universitaria
00100 Roma, ITALY

ZUNICA, L.
Departamento de Estadística
Facultad de Ciencias Economicas
y Empresariales
Avda. Blasco Ibáñez
Valencia-10, SPAIN

PREFACE

At conferences devoted to the foundations of probability and statistics, it is natural that attention should focus on points of division between supporters of rival schools of thought. The resulting confrontation of ideas and personalities in such contexts is often stimulating and useful in sharpening perceptions about one's own and other viewpoints.

But even at statistical meetings of a more general nature, a presentation or contribution from a statistician making positive use of Bayesian ideas all too often precipitates heated discussion about foundational ideas, with little or no attention directed to the detailed ideas or methods being put forward. This can be frustrating for those interested in specific theoretical or applied problems.

As in other areas of statistical discourse, the concentration on cleavage - to the exclusion of other features of interest- may be superficially entertaining, but it is not ultimately very productive.

And so part of our original motivation in organizing an International Meeting on Bayesian Statistics grew from the feeling that, although we are all loyal members of the good ship Statistics, there are times when a minority of the crew feel the need to head off into waters of their own choosing, unconstrained by existing charts and freed from the need to debate navigational philosophy before setting sail.

Of course, many voyages of discovery have already been undertaken. In particular, the semiannual NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics has become well established in the United States. And yet there has been little European participation in these meetings, a factor which led us to the more concrete idea of a conference held in Europe, attempting to draw together, for the first time at a truly international level, a fairly exhaustive assembly of statisticians concerned -in something other than a purely hostile sense! -with Bayesian statistics. Perhaps we also had in mind that this might accelerate the process of fertilization and growth of Bayesian ideas - a much needed acceleration if de Finetti's pessimistic prevision is to be confounded:

> My estimation is that another fifty years will be needed to overcome the present situation. It is based on the consideration that about thirty years were required for ideas born in Europe..... to begin to take root in America..... Supposing that the same amount of time might be required for them to

> establish themselves there, and then the same amount of time to return, we arrive at the year 2020.

(Theory of Probability, Vol. I, 1974, p. 2).

Whether or not this prophecy is affected by the proceedings of our meeting, it is clear from the list of participants that the need for an occasional forum of this kind is keenly felt, and the proceedings themselves provide a good overview of a wide range of current activity in Bayesian statistics, reflecting well the diversity of problems and viewpoints considered and expressed.

And yet a warm sense of unity prevailed at Hotel Las Fuentes; generated perhaps by that powerful mathematical current flowing from the wells of Tunbridge to the cold springs of Alcoceber - and celebrated in song at our final dinner:

"There's no Theorem like Bayes Theorem"
by G.E.P. Box

(to the tune of "There's no Business like Show Business")

VERSE(1)
The model, the data you can't wait to see
The theta, beta, sigma, and the rho
The Normal, the Poisson, the Cauchy, the t
The need to specify what you don't know
The likelihood for data you acquire
The perspicacious choosing of the prior

REFRAIN
There's no theorem like Bayes' theorem
Like no theorem we know

Everything about it is appealing
Everything about it is a wow

Let out all that a priori feeling
You've been concealing right up to now

There's no people like Bayes people
All odd balls from the urn

The other day you thought that you had got it straight
Take my advice and don't celebrate

A paradox by Lindley could arrive quite late
Another Stone to unturn

REFRAIN
There's no theorem like Bayes' theorem
Like no theorem we know

You can lose foreover that perplexed look
If you start to study it right now

Even more enthralling than a sex book
You'll find that textbook by Box and Tiao

There's no dogma like Bayes' dogma
Its great knowing you're right

We know of a fiducialist who knew the lot
We thought at first he had hit the spot

But after three more seminars we lost the plot
We just could not see the light

REFRAIN
There's no theorem like Bayes' theorem
Like no theorem we know

Fisher felt its use was quite restricted
Except in making family plans for mice

But there, he said, for pinning down a zygote
I'd give it my vote and not think twice

There're no answers like Bayes' answers
Transparent, clear and precise

Stein's conundrums you can solve without a blink
Best estimators in half a wink

You can even understand what makes 'em shrink
Their properties are so nice

VERSE(2)
There's Raiffa and Schlaifer, Mosteller & Pratt
There's Geisser, Zellner, Novick, Hill and Tiao
And these all are people who know what they're at
They represent Statistics' finest flower
And tho' on nothing else they could agree
With us they'd join and sing in harmony.

REFRAIN
There's no theorem like Bayes' theorem
Like no theorem we know

Just recall what Pearson said to Neyman
Emerging from a region of type B

"It's difficult explaining to the Lehmann
I fear it lacks Bayes' simplicity".

There's no haters like Bayes' haters
They spit when they see a prior

Be careful when you offer your posterior
They'll try to kick it right throught the door
But turn the other cheek if it is not too sore
Of error they may yet tire

REFRAIN *There's no theorem like Bayes' theorem*
Like no theorem we know

Critics carp at Bayes's hesitation
Claiming that his doubts on what he'd done
Led to late posthumous publication
We will explain that to everyone

When Bayes got up to Heaven
He asked for an interview
Jehovah quickly told him he had got it right
Bayes popped down earthwards at dead of night
His spectre ceded Richard Price the copyright
Its very strange but its true!

We would like to use this opportunity to thank the spanish authorities who made possible this Conference: specific mention should be made of the assistance provided by Prof. Cobo del Rosal and by Prof. Colomer. Finally, we are grateful to the staff of the Valencia University Press, and specially to Ms. Montse Blay and Ms. Ligia Saiz, who provided expert assistance in typing and proofing the manuscript.

J.M. Bernardo
M.H. DeGroot
D.V. Lindley
A.F.M. Smith

Trab. Estadist. (1980)
Vol. 31 pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith eds.
Bayesian Statistics (1980)
pp. 15-66 Valencia, Spain: University Press

1. Foundations of Subjective Probability and Decision Making

INVITED PAPERS

GIRON, F.J. and RIOS, S. (*Universidad de Málaga, Universidad de Madrid*)
Quasi-Bayesian Behaviour: a more realistic approach to decision making ?

HILL, B.M. (*University of Michigan and University of Utah*)
On finite additivity, non-conglomerability, and statistical paradoxes

DISCUSSANTS

GOOD, I.J. (*Virginia Polytechnic and State University*)
PICCINATO, L. (*Università di Roma*)
VILLEGAS, C. (*Simon Fraser University*)
DICKEY, J.M. (*University College of Wales*)
DEGROOT, M.H. (*Carnegie-Mellon University*)
FRASER, D.A.S. (*University of Toronto*)
FRENCH, S. (*University of Manchester*)
LINDLEY, D.V. (*University College London*)

REPLY TO THE DISCUSSION

Quasi-Bayesian Behaviour: A more realistic approach to decision making?

F.J. GIRON
Universidad de Malaga
and
S. RIOS
Universidad de Madrid

SUMMARY

In this paper the theoretical and practical implications of dropping-from the basic Bayesian coherence principles- the assumption of comparability of every pair of acts is examined. The resulting theory is shown to be still perfectly coherent and has Bayesian theory as a particular case. In particular we question the need of weakening or ruling out some of the axioms that constitute the coherence principles; what are their practical implications; how this drive to the notion of partial information or partial uncertainty in a certain sense; how this partial information is combined with sample information and how this relates to Bayesian methods. We also point out the relation of this approach to rational behaviour with the more (and apparently unrelated) general notion of domination structures as applied to multicriteria decision making.

Keywords: COHERENCE PRINCIPLES; AXIOMS OF RATIONALITY; PARTIAL UNCERTAINTY; BAYESIAN PREORDERS; PARTIALLY ORDERED PROBABILITIES; UPPER AND LOWER PROBABILITIES; DOMINATION STRUCTURES.

1. INTRODUCTION

As it is well known, Bayesian coherence principles as applied to decision problems imply the existence of a utility function, unique up to a linear transformation, and what is more important from the inferential point of view, a *unique* probability measure (known as subjective or personal probability) such that in order to choose among acts that which maximizes expected utility is selected.

Thus these principles assume that in any decision problem under uncertainty the decision maker is able - by instrospection or by any other means - to assign unique probabilities to every possible event and he will choose the decision which maximizes his expected utility.

On the other hand, if nothing is known about the true state of Nature, and one does not want to stick to incoherent principles such as minimax, etc., the only solution is to turn one's attention to admissible acts or decisions by application of the dominance principle implied by the natural ordering of decision rules once utilities have been assigned.

Between these two cases: 1st) - prior distribution is completely known; and 2nd) - nothing is known about the prior distribution (except the trivial fact that it is a probability measure, the existence of which may be even questioned), we may place the case of *partial ignorance* or *partial uncertainty*.

What we call partial ignorance refers to the fact we represent our knowledge about states of Nature by means of a set of probability measures to which the true distributions belongs.

In a more general sense "partial ignorance" could represent information about the states of Nature **not necessarily** given in the form of probability distributions. However our axioms or rationality principles will rule out this second interpretation of partial ignorance. In other words, we shall prove that a weakening of Bayesian coherence principles characterize partial ignorance in terms of a set of probability measures and that this characterization embodies the two extreme cases (of total ignorance, and perfect knowledge of prior distribution) which are but particular cases.

The idea of representing partial ignorance by convex sets of probability measures or by means of the related concept of lower and upper probabilities is not new and dates back to Smith (1961, 1965), Good (1962) and Dempster (1968)[1], and more recently to Suppes (1974) and Ríos (1975a, 1975b, 1976). However, none of these authors give a complete characterization of partial ignorance. Smith (1961) gives a partial answer to this question for the finite case. More refined results are found in Girón (1978).

Partial ignorance may be looked at in two different ways. First, suppose the decision maker is uncertain about his *prior P* so he expresses his beliefs in the form of a statement such as *P belongs to* K^*. The form and size of K^* measures his relative uncertainty. It is remarkable (see theorems 3.2, 3.4 and 3.6) that if a decision maker reflects his uncertainty about states of Nature in such a way that he is not able to compare every pair of acts (Axiom A1, sec. 3) whilst other axioms hold, then his uncertainty can be measured in terms of a

[1] As early as 1940, Koopman (1940) pioneered the idea that not every pair of events are comparable. In our approach this is a result of dropping the completeness axiom *C* (see section 3).

set of probability measures and he compares acts in terms of expected utilities against the probability measures of this set.

Second interpretation runs as follows: suppose an arbitrary number of decision makers each one being perfectly coherent (that is, their preferences satisfy axioms A1 to A5, and C of section 3). Suppose further that the utilities they assign to consequences are in agreement but they differ in their preferences, that is, their personal or subjective probabilities differ. Then the intersection of their preference systems is a new preference relation satisfying axioms A1 to A5. In this case, partial information or uncertainty is represented by the convex set generated by the set of all prior distributions corresponding to the decision makers, which, in this second version, could be named the **feasible set.**

If we call coherence principles the axioms A1 to A5, and C, we shall now discuss, briefly, the implications of dropping any of them. We do not discuss the necessity of axioms A1, A2, A3 and A4 as it is well known from the literature that dispensing with any of them drive to incoherent decisions.

As to axiom A5 we could dispense with it. In this case the preference relation would be a lexicographic order that would be characterized by a multidimensional (or lexicographic) subjective probability $P = (P_1, P_2, \ldots)$.

So the principle under discussion is completeness (axiom C). In its favour one may say that whichever the decision or inferential problem one is faced at a decision has to be made, and this imply that the decision maker or statistician is able to compare every pair of acts. However in case of partial ignorance the decision maker restricts his attention to non-dominated decisions. If this set is a small one and the corresponding Bayes risks do not differ much, this might be considered as though one would be performing a sensitivity analysis in a Bayesian case (e.g., see Fishburn (1964)).

From the purely inferential view point both approaches - partial versus Bayesian knowledge - are even closer. In the Bayesian case all information is in the posterior distribution while in the quasi-Bayesian case all relevant information is in the posterior set. But this last situation can be assimilated to the first one by taking a greater sample (see, e.g., example a) of section 4).

Note the difference between dropping the completeness axiom in utility theory (Aumann (1962, 1964), Criado (1978)) and in subjective probability theory. In the first case partial knowledge of utility function is not reduced (in fact sample information is independent of utility) by sample information; yet in case of partial knowledge of prior distribution, sample information reduces uncertainty[2]. That means that our initial partial preorder converges to a

[2] See Girón (1979) for a discussion on duality between the concepts of utility and subjective probability.

complete preorder when sample size increases.

Thus dropping completeness axiom is not made for sake of mathematical generalization but to convey a rational model for the case when it is difficult to choose among decisions. The practical conclusion is: "if you feel unsure about your decisions, then take a greater sample than the one you would take if you were able to compare every pair of decisions and you will do (nearly) as well".

2. DECISION MODEL WITH PARTIAL INFORMATION

Let $(\Omega, D; L)$ be a decision problem, where Ω is a set of states of Nature or parameter space that for illustrative purposes we suppose is finite and will be denoted $\Omega = \{\theta_1,\dots,\theta_n\}$ (later on this section this restriction will be lifted); D is a set of possible decisions, which allowing for randomization may be supposed convex, and L is a loss function (the negative of a utility function), that is:

$$L: \Omega x D \rightarrow \mathbb{R}.$$

In the Bayesian case we also have information on Ω given in the form of a single probability measure, known as "the prior distribution", which we denote by P. In our case P can be identified with a point of the n-simplex of R^n that will be denoted

$$\Omega^* = \{ (p_1,\dots,p_n); \Sigma_{i=1}^{n} p_i = 1;\ p_i \geq 0; i = 1,\dots,n\},$$

where it is understood that $p_i = P(\theta_i)$, so that Ω^* would be the set of all probability measures.

If K^* is a nonempty subset of Ω^*, then partial information about P (the "true" prior distribution) is to state simply that $P \in K^*$. If K^* in fact represents partial ignorance, it may be taken to be convex, for if the decision maker is uncertain about P_1 and $P_2 \in K^*$, then he is uncertain about $\alpha P_1 + (1\text{-}\alpha) P_2$ $(0 \leq \alpha \leq 1)$. So convexity of K^* is not introduced for mathematical convenience but as a fairly natural condition.[3]

Def. 2.1. A decision model with partial information is a quadruplet $(\Omega, D; L; K^*)$ where K^* is a nonempty closed convex set of Ω^*, which will be called the **uncertainty set** or the **prior distribution set.**

As particular cases we have: 1) the case of complete ignorance, when $K^* = \Omega^*$; 2) the case of perfect knowledge of the prior distribution when $K^* = \{P\}$, that is, K^* reduces to a point, or Bayesian case.

As most of the ideas we are to set forth have simple geometrical interpretations, it will be convenient to transform the decision problem into an equivalent S-game[4] as follows:

Define the risk set S of decision problem $(\Omega, K; L)$ by

$$S = \{x = (x_1,\dots,x_n); \exists\, d \in D;\ L(\theta_i, d) = x_i\}$$

Let us consider the simplest case of two states of Nature, that is, $\Omega = \{\theta_1, \theta_2\}$. Then the partial information about $P = (p_1, p_2)$ is given in its more general form, by inequalities

$$\alpha_1 \leq p_1 \leq \alpha_1',$$

with α_1, α_1' constants such that $0 \leq \alpha_1 \leq \alpha_1' \leq 1$.

The set K^* can be geometrically represented by the angle determined by the extreme vectors $(\alpha_1, 1\text{-}\ \alpha_1)$, $(\alpha_1', 1\text{-}\alpha_1')$. Let $x^* = (x_1^*, x_2^*)$ be a fixed point of the risk set. Then the Bayes risk for x^* against prior distribution $P = (p_1, p_2) \in K^*$ is

$$r(x^*; P) = p_1 x_1^* + p_2 x_2^*.$$

If we take as priors the extreme point of K^*, say $P_1 = (\alpha_1, 1\text{-}\alpha_1)$ and $P_1' = (\alpha_1', 1\text{-}\alpha_1')$, the corresponding Bayes risks are

$$r\,(x^*; P_1) = \alpha_1 x_1^* + (1\text{-}\alpha_1) x_2^*$$
$$r(x^*; P_1') = \alpha_1' x_1' + (1\text{-}\alpha_1') x_2^*,$$

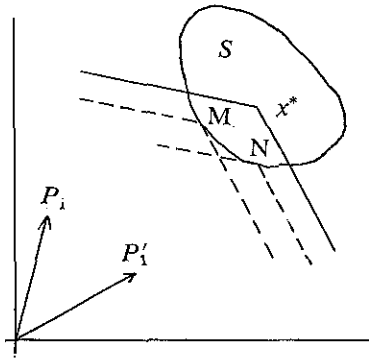
If we now consider the set of (possible) points that have smaller Bayes risk against both P_1 and P_1', we see that these points lie in the intersection of the closed half-planes given by the following inequalities

[3] The topological condition of K^* being closed is not really essential for as we shall show either K^* or $\overline{K^*}$ (its closure) generate the same quasi-Bayesian preorder. Note that in the Bayesian case, K^* reduces to a point which is closed. Convexity could also be dispensed with as it can be shown that K^* and con (K^*) (convex hull) generates the **same quasi-Bayesian preorder.**

[4] For a definition of S-games see Blackwell and Girshick (1954), that suffices for the finite case. For a more general definition see Girón (1975).

$$\alpha_1 x_1 + (1-\alpha_1)\, x_2 \leq r\,(x^*, P_1)$$
$$\alpha_1' x_1 + (1-\alpha_1')\, x_2 \leq r\,(x^*, P_1') \qquad (2.1)$$

that define an angle with vertex at x^* (see figure). The most



important point to notice is that every point belonging to the angle, that is, satisfying inequalities (2.1), has smaller Bayes risk than x^* **against any** prior distribution $P \epsilon K^*$.

A further point to notice is that the angle does not depend on the chosen x^*: that is, if $y^* \neq x^*$ then the angle corresponding to y^* is simply a translation of the one with vertex at x^*. We shall denote this angle with vertex at origin by K. So K depends only on K^*.

This itself suggests the idea of defining a partial preorder on S (which is extended to R in an obvious way) by means of angle K and then, regard as solutions of the decision problem the maximal points (decisions) in S (in D). Thus maximal points in this weak order will coincide[5] with Bayes solutions against all probability measures belonging to K^*.

In the above figure this set is represented by the arch MN.

Note that in the case of total ignorance, that is, $\alpha_1 = 0$, $\alpha_1' = 1$, the angle defined by (2.1) is precisely the set of points $x = (x_1, x_2)$ such that

$$x_1 \leq x_1^*, x_2 \leq x_2^*;$$

[5] This is not true as stated because the relation between maximal and Bayes solutions in this case is analogous to the existing relation in the well-known case of natural ordering. For details see Rios (1976).

that is, the natural ordering of risk points in R^n.

The last and most important point to notice is the relation existing between K and K^*. In the simple case considered K is but the polar cone of set K^*. Thus partial knowledge represented by K^* induces in the space of possible decision functions a "domination structure" which is characterized by the polar cone of set K^*.

Recall (see definition of polar cone below) that the polar cone K is closed and that polar cones of K^* and $\overline{K^*}$ are the same. Further in the example considered the polar cone of K^* and of the set of its extremal points $\{P_1, P_2\}$ is the same.

These mathematical properties justify the hypotheses put on the set K^* of convexity and closedness. In next section these properties will be justified, through an axiomatic approach, from simple coherence principles.

Let us now return to the case of a finite number of states of nature $\{\theta_1, \ldots, \theta_n\}$. We are to define what we understand by quasi-Bayesian preference relations associated to a decision problem with partial ignorance.

Def. 2.2. Let $(\Omega, S; K^*)$ be a decision problem with partial uncertainty. We shall call K^***-Bayesian preference** or **quasi-Bayesian preference** in S to the relation $\gtrsim_k^*$ defined for every $x, y \epsilon S$ by

$$x \gtrsim_k^* y \text{ if and only if } x.P \leq y.P \text{ for every } P \epsilon K^*,$$

where $x.P$ denotes dot-product.

It can be shown that $\gtrsim_k^*$ is a weak partial order satisfying axioms A1 to A5 of section 3. Moreover, $\gtrsim_k^*$ is complete (or linear if and only if K^* reduces to a point $\{P\}$: In this last case $\gtrsim_p$ is called a **Bayesian preference relation.**

Let K^* be the uncertainty set; denote by K the polar cone of set K^*, that is

$$K = \{x = (x_1, \ldots, x_n) \epsilon \mathbb{R}^n ; x.P \leq 0 \text{ every } P \epsilon K^*\}.$$

K is a closed convex cone of $\mathbb{R}^n$ with vertex at origin. This defines a preference relation in $\mathbb{R}^n$ (and consequently in S) as follows

Def. 2.3. Let $x, y \epsilon S$. x K- **dominates** y and will be denoted $x \geq_k y$ if and only if $x\text{-}y \epsilon K$.

The relation between the two definitions, which is but a consequence of duality, is the content of next result.

Theorem 2.1.

$$x \gtrsim_k^* y \text{ if and only if } x \gtrsim_k y.$$

It is worth mentioning that if K reduces to a point P, then K is the closed half-space defined by $\{x \epsilon \mathbb{R}^n \ ; \ x \bullet P \leq 0\}$. In case of total ignorance K is the negative orthant $D_o = \{x \epsilon \mathbb{R}^n, x_i \leq 0, i = 1,2,,\ldots,n\}$.

Def. 2.4. Let $\gtrsim$ and $\gtrsim^*$ be two weak order relations. Then, relation $\gtrsim$ **is included in** $\gtrsim^*$ if and only if $x \gtrsim^* y$ implies $x \gtrsim y$.

Theorem 2.2. Let K_1^*, and K_2^* be subsets of Ω^*, then $K^*_1 \supset K_2^*$ implies $\gtrsim_{k1}{}^*$ is included in $\gtrsim_{k2}^*$. Moreover, if K_1^*, and K_2^* are closed convex sets, the conversely statement is also true.

As a consequence of duality and theorem 2.1., we have the following

Corollary 2.1. $\gtrsim_{k1}$ is included in $\gtrsim_{k2}$ if and only if $K_1 \subset K_2$.

These partial weak orderings give rise to definitions of admissibility, complete classes and quasi-Bayesian (or K-Bayesian) decisions.

Def. 2.5. A risk point $x \epsilon S$ is **K^*-Bayes (or quasi-Bayesian)** for the problem (Ω, S; K^*) if there exists at least a $P \epsilon K^*$ such that x is Bayes against P. Accordingly $d \epsilon D$ is **K^*-Bayes** if its corresponding risk point is K^*- Bayes.

The set of all K^*-Bayes strategies will be denoted $\mathbf{B}(K^*; S)$ or $\mathbf{B}(K^*; D)$.

Relations among K-admissibility (defined in an obvious way), K-Bayesness and completeness can be found in Ríos (1976), in which the name "quasi-Bayes" was coined.

In this paper we do not discuss the computational aspects of quasi-Bayesian strategies. In the finite case, here considered, methods for finding non-dominated strategies are to be found in Leitmann (1976) and references therein. These procedures, devised for general convex domination structures, can be applied *mutatis mutandis* to the problem of finding quasi-Bayesian strategies in case K^* be a convex polyhedron by means of linear and non-linear programming technics. The general case of K^* being an arbitrary convex set may be treated by approximative methods (see reference above).

By far the most important feature of quasi-Bayesian methods is that they allow incorporation of the information provided by an experiment by use of Bayes theorem.

Let (X, A_z, P_θ (x)) be an experiment, where $\Omega = \{\theta_1,\ldots,\theta_n\}$. Let $P(\theta_i|x)$ denote the posterior probability of θ_i when x has been observed and prior is $P(\theta_i)$. We define the **posterior uncertainty set** (or posterior partial information set) as the set of all posterior distributions of K^* when x is observed. This set will be represented by K^*_x. Sometimes we shall refer to this set K_x^* **as the transform of K^* through sampling when x is observed.** Properties of posterior uncertainty sets are summarized in the following.

Theorem 2.3. If K^* is a closed convex set of Ω^*, then K^*_x is also a closed convex set for every $x \epsilon X$. Furthermore, extremal prior distributions of K^* are transformed through sampling into extremal distributions of K^*_x for any $x \epsilon X$.

The second part of theorem usually simplifies the problem of finding the posterior uncertainty set if only we know the extremal prior distributions.

Finally, we mention the fact that the whole set of probability distributions Ω^* is invariant through sampling, that is $\Omega^*_x = \Omega^*$ for any $x \epsilon X$. This is but a statement that total ignorance cannot be changed into partial ignorance through sampling.

Def. 2.6. Let (Ω, D; L; K^*) be a decision problem with partial information, (X, A_x; $P_\theta(x)$) an experiment. We say δ: (X, A_x) → (D, A_D) is K^*-Bayes (or quasi-Bayes) if for every $x \epsilon X$, $\delta(x)$ is K^*-Bayes for the problem (Ω, D; K_x^*).

Most definitions and results given in this sections are easily generalizable to the case of an infinite number of states with slight modifications except in one instance. This refers to the duality between K^* and its polar cone K that poses delicate analytical problems due, in part, to the lack of reflexivity (in the sense of functional analysis) of some of the spaces of measures under consideration, and secondly to the problem that appears in some statistical applications that D and Ω^* cannot be embeded in topological vector spaces for which one is the dual of the other one.

3. AXIOMATIC CHARACTERIZATION OF PARTIAL UNCERTAINTY

As we stated in the introduction, partial uncertainty is usually represented by a convex set of probability measures and may be considered midway between total ignorance (no knowledge of the "true" [if any] prior distribution) and, on the other hand, whole knowledge of the prior distribution (Bayesian view point).

Most axiomatic characterizations of subjective probability and, consequently, of Bayesian behaviour in the case of decisions under uncertainty are based in the ability of the decision-maker at ordering **any** pair of acts or events he is confronted with; which, as it is well known, is one of the basic principles of the so called "Bayesian coherence principles".

Here we present a variant of the above mentioned principles that still preserve the Bayesian "flavour" but have into account this possibility and, in fact, they fully characterize "partial ignorance". Basically, we follow the axioms given by Girón (1974, 1977)[6] that characterize subjective probability and the principle of maximization of expected utility.

The basic idea is the suppression of the completeness of the preference relation in the set of all possible decisions, along the lines of Aumann's contribution to utility theory [see, Aumann (1962, 1964)], which could justify the name of "subjective probability without the completeness axiom" instead of "partial ignorance".

One of the main results of this section is the characterization of **all** partial ignorance relations (this includes the Bayesian case) in terms of a class of closed convex cones in the space of decisions. This first characterization is inspired in the papers of S. Ríos (1975a, 1975b, 1976) on quasi-Bayes orders, and, on the other hand, in the work of Yu, Zeleny *et al.*[7] on domination structures.

The second characterization is the basic result we are seeking for; stated in imprecise terms it asserts that partial ignorance is characterized in terms of closed convex subsets of a space of probability measures.

In the following it will be convenient to distinguish two cases; namely: a) partial ignorance is represented in terms of σ-additive probability measures (abbreviated p.m.); b) these probabilities are only assumed to be finitely additive.

In case a) (See, e.g. Girón (1977), p. 33) a restriction on the set of states of Nature needs be imposed; namely, it is supposed to be a compact Hausdorff topological space; a further restriction is that decisions can be identified with a subset of continuous functions on such a space. However, in case b), the parametric space can be quite arbitrary and decisions or acts are only supposed to be bounded.

Case a), in spite of its apparent restrictiveness, it is not so, for many decision problems are such that the parameter space may be endowed with a metric (e.g., the intrinsic metric) which makes all acts continuous so that the only restriction would now be compactness respect to that topology.

Let Ω denote the space of states of Nature or parameter space, D a set of decisions on terminal acts, and u: $\Omega x D \to \mathbb{R}$ a utility function.

Def. 3.1. A **decision problem under uncertainty** (which, in the sequel, will be abbreviated as d.p.u.u.) is a triplet $(\Omega,D;u)$.

In case a) D can be identified with a subset S of C (Ω)- space of all real continuous functions defined on Ω -namely

[6] These axioms were put forward in a later paper (see Girón (1978)), in a stronger form that the one given in this article.

[7] Most of their work appears in Leitmann. (1976).

$$S = \{f\,(\theta)\epsilon\; C(\Omega);\; \exists\; d\epsilon D; f\,(\theta) = u\;(\theta,d)\}.$$

In case b) D is identified with a subset S of B (Ω) -space of all real bounded functions defined on Ω -defined by

$$S = \{f\,(\theta)\epsilon B\;(\Omega);\; \exists\; d\epsilon D; f\,(\theta) = u(\theta,d)\}.$$

Further, if the decision maker or statistician allows for randomization in D, S may be regarded as a convex subset of the linear spaces, $C(\Omega)$ and $B(\Omega)$, respectively.

This suggests a new definition of a d.p.u.u.

Def. 3.2. A d.p.u.u. is a pair (Ω,S) where S is a nonempty convex subset of C (Ω) (case a) or B (Ω) (case b).

Def. 3.3. If (Ω,S) is a d.p.u.u. a **decision criterion** is a binary relation on S, which will be denoted by $\gtrsim_s$

Relation $\gtrsim_s$ is read "...at least as preferred as...". Taking $\gtrsim_s$ as the basic relation we may define the following.

$$f \gtrsim_s g \text{ iff } f \gtrsim_s g \text{ and not } g \gtrsim_s f$$

$$f \sim_s g \text{ iff } f \gtrsim_s g \text{ and } g \gtrsim_s f$$

$$f \not\curlywedge_s g \text{ iff not } f \gtrsim_s g \text{ and not } g \gtrsim_s f$$

which are read "...(strictly) preferred to...", "...indifferent to..." and "...is not comparable to...", respectively. In the sequel $g \lesssim_s f$ will mean $f \gtrsim_s g$.

The list of proposed axioms is the following, that only differs of Girón's (1977) in the first one.

A1 (Partial preorder).- For every d.p.u.u. (Ω,S), $\gtrsim_s$ is reflexive and transitive.

A2 (Strong dominance).- If f, $g \epsilon S$ are such that $f\,(\theta) > g\,(\theta)$ for every $\theta \epsilon \Omega$, then $f >_s g$.

A3 (Addition of new strategies).- If $S \subset R$, then $f \gtrsim_s g$ implies $f \gtrsim_R g$.

A4 (Linearity).- If $\lambda \epsilon$ $(0,1)$, $f,g,h \epsilon S$, then $f \gtrsim_s g$ if and only if

$$\lambda f + (1\text{-}\lambda)\, h \gtrsim_s \lambda g + (1\text{-}\lambda) h$$

A5 (Continuity).- If f_n, g, $h \epsilon S$ for $n = 1,2,\ldots$, are such that $\{f_n\} \to f \epsilon S$,

$$f_n \gtrsim_s g,\; h \gtrsim_s f_n$$

for every $n = 1,2,\ldots$, then $f \gtrsim_s g$ and $h \gtrsim_s f$.

Convergence in this axiom is understood with respect to the usual supremum norm topology given, for both a) and b) cases,

$$\| f \| = \sup_{\theta \epsilon \Omega} |f(\theta)|.$$

Next is a completeness axiom that will only be necessary in the characterization of Bayesian behaviour.

C (Completeness). For every $f, g \epsilon S$ either $f \gtrsim_s g$ or $g \gtrsim_s f$

Axiom 3 allows us to consider the $\gtrsim$ relation as being defined on $C(\Omega)$ $[B(\Omega)]$; then relation $\gtrsim_s$ is, simply, the restriction of $\gtrsim$ to S. Further, as $C(\Omega)$ $[B(\Omega)]$ are complete normed spaces, if $\{f_n\}$ converges to f, then $f \epsilon C(\Omega)$ $[B(\Omega)]$.

In case b), as $B(\Omega)$ contains the class of indicator functions of subsets of Ω, the relation $\gtrsim$ on $B(\Omega)$ restricted to this class allows us to define a new relation, $\gtrsim^*$, on the class of all subsets of Ω, which we shall denote by $P(\Omega)$, and will be called events.

Def. 3.4. Event A is **at least as probable as event B,** and will be denoted by $A \gtrsim^* B$ if $\lambda > \mu$ implies $f \gtrsim g$, where f and g are defined by

$$f(\theta) = \begin{cases} \lambda & \text{if } \theta \epsilon A \\ \mu & \text{if } \theta \notin A \end{cases}$$

$$g(\theta) = \begin{cases} \lambda & \text{if } \theta \epsilon B \\ \mu & \text{if } \theta \notin B \end{cases}$$

It can be easily seen that if $\gtrsim$ satisfies axioms A1, A2, A3, A4, A5, definition 3.4. does not depend on λ and μ, as far as $\lambda > \mu$. This is the content of the following lemma, which could have been taken as definition.

Lema 3.1.- A $\gtrsim^* B$ if and only if $I_A \gtrsim I_B$, where I_A and I_B denote the indicator functions of sets A and B, respectively.

Furthermore, relation $\gtrsim^*$ as defined above satisfies all axioms of comparative probability (e.g., see Fine (1973) p. 17) except the comparability of every pair of acts[8]. In particular $\gtrsim^*$ satisfies

(i) $\gtrsim^*$ is reflexive and transitive.

(ii) $A \gtrsim^* \phi$ for every event $A \epsilon P(\Omega)$.

(iii) $\Omega \gtrsim^* \phi$

(iv) Let A,B,C be events such that $A \cap C = B \cap C = \phi$, then

$$A \gtrsim^* B \text{ if and only if } A \cup C \gtrsim^* BUC.$$

As was mentioned in section 2, the natural ordering in $C(\Omega)$ $[B(\Omega)]$ is the weakest partial preorder every other "reasonable" partial, or complete, preorder should be consistent with. This consistency is taken up in the formulation of axioms A2 and A5.

Def. 3.5. f **dominates** g, and will be represented by $f \gtrsim_d g$ if $f(\theta) \geq g(\theta)$ for every $\theta \epsilon \Omega$.

Relation $\gtrsim_d$ is a partial preordering satisfying axioms A1, A2, A3, A4, A5. Morover, relation $\gtrsim^*_d$ induced in $P(\Omega)$ by $\gtrsim_d$ is subset inclusion, e.i.,

$$A \gtrsim_d B \text{ if and only if } A \supseteq B.$$

Those decisions dominated by the function $I_\phi \equiv o$ will be denoted by D_o, that is,

$$D_o = \{f;\ f(\theta) \leq 0 \text{ for every } \theta \epsilon \Omega\}$$

Some of the results that now follow were advanced in Girón (1978)[9].

Theorem 3.1. If relation $\gtrsim$ in $C(\Omega)$ $[B(\Omega)]$ satisfies A1, A2, A3, A4, A5 then there exists a unique closed convex cone K, K not being the entire space, containing D_o and with vertex at the origin, such that

$$f \gtrsim g \text{ if and only if } g\text{-}f \epsilon K \tag{3.1}$$

Conversely, every non empty closed convex K, containing D_o and with vertex at the origin, defines a partial preordering $\gtrsim$ in $C(\Omega)$ $[B(\Omega)]$ by (3.1)

Furthermore, $\gtrsim$ is a complete preordering if and only if K is a closed half-space containing D_o and passing through the origin.

[8] Recently Fishburn (1975) and Goodman (1977) have also considered a weakening of the comparability axiom in which indifference is not assumed to be transitive.

[9] In this paper we give new results and some refinements and amendments of results that appeared in Giron (1978). Proofs will appear in a subsequent paper.

This theorem is interesting in order to examine the structure of partial preorderings in relation to complete preorders.

Let $\gtrsim_i$ be a collection of linear preorders satisfying axioms A1, A2, A3, A4, A5, and C, where $i\epsilon I$, a certain index set. If we define relation $\gtrsim$ by

$$f \gtrsim g \text{ if and only if } f \gtrsim_i g \text{ for every } i\epsilon I,$$

then $\gtrsim$ is a partial preordering satisfying A1, A2, A3, A4, A5. This relation could be named the **intersection of the class of preorderings** $\{\gtrsim_i\}i\epsilon I$

Now, by theorem 3.1, every partial preorder is characterized by a closed convex cone K and every complete preorder by a closed half-space, so that we have as a corollary of the theorem the following.

Corollary 3.1. Every partial preorder satistying A1, A2, A3, A4, A5 is the intersection of an arbitrary collection of linear preorderings satisfying A1, A2, A3, A4, A5 and C, and conversely.

It can also be shown that the intersection of an arbitrary collection of partial preorders satisfying A1 to A5 is a partial order satisfying A1 to A5.

If we call "quasi-Bayesian preorder" then corollary 3.1 simply states that every "quasi-Bayesian preorder" is the intersection of Bayesian preorders, thus giving a precise meaning to the second interpretation of partial ignorance mentioned in the introduction.

Next theorem, and its counterpart for case b) (see theorem 3.4), characterizes a partial ignorance in terms of a set of probability measures.

Theorem 3.2. If relation $\gtrsim$ in $C\,(\Omega)$ satisfies A1 to A5 then there exists a unique non empty closed convex set K^* of σ-additive probability measures on the Borel field of the topological space (Ω,B_Ω) such that

$$f \gtrsim g \text{ if and only if } \int f\, d\mu \geq \int g d\mu \text{ for every } \mu\epsilon K^*$$

If $\gtrsim$ further satisfies axiom C, then K^* reduces to a single probability measure.

The second part of this theorem characterizes Bayes behaviour.

Technical note. In this theorem as well as in theorem 3.4 below, K^* is closed in the weak * topology.

Next theorem characterizes the natural ordering relation $\gtrsim$ in case a), the necessary part of the theorem being as well known result in integration theory. In fact, it is a particular case of theorem 3.2 that characterizes total ignorance.

Theorem 3.3. For every $f, g \in C\,(\Omega)$

$$f \gtrsim g \text{ if and only if } \int f d\mu \geq \int g d\mu$$

for every $\mu\epsilon\Omega^*$, where Ω^* is the set of all probability measures (σ-additive) on the space (Ω,B_Ω).

The corresponding theorems for case b) are:

Theorem 3.4. If relation $\gtrsim$ on $B\,(\Omega)$ satisfies A1 to A5, then there exists a unique nonempty closed convex set K^* of finitely additive probability measures on $P\,(\Omega)$ such that

$$f \gtrsim g \text{ if and only if } \int f dP \geq \int gdP \text{ for every } P\epsilon K^*$$

If $\gtrsim$ further satisfies axiom C, then K^* reduces to a unique probability measure.

Theorem 3.5. For every $f, g\epsilon B\,(\Omega)$

$$f \gtrsim_d g \text{ if and only if } \int f dP \geq \int gdP$$

for every $P\epsilon\Omega^*$, where Ω^* is the set of all finitely additive probability measures on the space $(\Omega,P\,(\Omega))$.

Next two theorems refer to the comparative probability relation $\gtrsim^*$ of definition 3.4 or lemma 4.1.

Theorem 3.6. A $\gtrsim^* B$ if and only if $P\,(A) \geq P\,(B)$ for every $P\,\epsilon K^*$, where K^* is the set of theorem 3.4.

Theorem 3.7. For every pair of events $A,B\,\epsilon P\,(\Omega)$

$$A \supseteq B \text{ if and only if } P\,(A) \geq P\,(B)$$

for every $P\,\epsilon K^*$, where K^* is the set defined in theorem 3.5.

Theorem 3.6 could be used to define a system of lower and upper probabilities associated to the CP partial relation $\gtrsim$, in the following manner

$$P_*\,(A) = \operatorname*{Inf}_{P\,\epsilon K^*} P\,(A),$$

$$P_*\,(A) = \operatorname*{Sup}_{P\,\epsilon K^*} P\,(A)\,.$$

Yet the properties of P_*, P^* will not be further explored in this paper, as our intention was to fully characterize partial ignorance.

This section ends with a few results referring to conditional preference. They essentially show that the intuitive ideas set forth in section 2 about the incorporation of information given by an experiment to partial prior ignorance, given in the form of a convex set of probability measures, through the use of Bayes theorem are sound and have an axiomatic foundation. It is also proven that the posterior set of probability measures is also a closed convex set, which generalizes last theorem of section 2.

Definition of conditional preference appears in a different form that the one given in Savage (1954) and Girón (1977) for the sake of mathematical tractability.

Def. 3.6. Let f and g be two given acts. f **is at least as preferred as** g **when** A **obtains**, and will denoted $f \gtrsim g$ **given** A, if and only if $I_A f \gtrsim I_A g$.

Def. 3.7. Event A **is null**, if and only if $f(\theta) > g(\theta)$ for every $\theta \in \Omega$ does not imply $f \gtrsim g$ given A.

Properties of null events derived from axioms A1 to A5 are similar to the ones given by Savage (1954). In particular we have

(i) ϕ is a null event.
(ii) If A is null and $B \in A$, then B is null.
(iii) The union of any finite number of null events is null.
(iv) Ω is not null.

In terms of the set K^* null events are characterized by the following:

Theorem 3.8. A is null if and only if there exists at least a $P \in K^*$ such that $P(A) = o$.

Next lemma is a trivial consequence of definition 3.7., but conveys an important result in conjunction with theorem 3.4.

Lemma 3.2. If A is not null, relation $\lesssim$ given A, satisfies axiom A1 to A5.

Next theorem characterizes conditional preference.

Theorem 3.9. If axioms A1 to A5 hold and event A is not null, then there exists a unique closed convex set $K_A^* \subset \Omega^*$ such that

$$f \gtrsim g \text{ given } A, \text{ if and only if, } \int f\,dP \geq \int g\,dP$$

for every $P \in K_A^*$.

The relation between sets K^* and K_A^* of theorems 3.4 and 3.9 is given by the following theorem that shows that K_A^* is precisely the set of all conditional probability measures of K^*.

Theorem 3.10. If A is not null, then

$$K_A^* = \{P_A \epsilon \Omega^* ; \exists\, P \epsilon K^* ; P_A(B) = \frac{P(A \cap B)}{P(A)} \text{ for every } B \epsilon P(\Omega)\}$$

This has a clear behavioural interpretation in terms of intersection of orders: We know from theorems 3.4, 3.6 and corollary 3.1, that every quasi-Bayesian preorder is the intersection of quasi-Bayesian preorders. Now, suppose we are given the piece of information that "event A has obtained" and A is not null. It can be easily shown that if the partial preorder $\gtrsim$ is the intersection of $\gtrsim_i$, for $i \epsilon I$, then A is not null for $\gtrsim_i$ for every $i \epsilon I$. If $\gtrsim_i$ is characterized by subjective probability P_i and event A obtains, then P_i is changed into P_{iA} to which corresponds $\gtrsim_i$ given A, so that $\gtrsim$ given A is precisely the intersection of the $\{\gtrsim_i$ given $A\}_{i \epsilon I}$. This is in the spirit of Bayesian behaviour: «Change your prior partial information through use of Bayes theorem into the posterior partial information and act accordingly to the principle given in theorems 3.2 and 3.4 which could be named the **principle of maximalization of expected utility**».

As was pointed out at the end of section 2 partial ignorance can be characterized by the extreme point of set K^*, for as if we denote it by K^*_e, then $K^* = \overline{\text{con}}\ (K_e^*)$, so that any possible distribution is a general mixture of extreme distributions. It can be easily shown that extremal prior distributions change into extremal posterior distributions by use of Bayes theorem.

4. ILLUSTRATIVE EXAMPLES

In the last section we give a few simple examples in order to illustrate the form of quasi-Bayesian solutions.

In case quasi-Bayesian procedures are intended only for inferential purposes the answer lies on the structure of the posterior set of probability measures, or to reduce it to a minimum, all relevant information is given by the set of extremal distribution of this set.

In the case of decision problems, a loss or utility structure is imposed upon the inferential problem, thus reducing the decision problem to the calculation of a few parameters of the posterior extremal distributions, those parameters depending on the form of the loss function.

a) Quasi-Bayesian confidence intervals in the normal case

Suppose $X_1,\ldots,X_n$ is a random sample of a normal distribution $N(w,r)$ where the precision r is known and the mean w is unknown. The partial information on w is given by the subset of normal distributions $N(\mu,\tau)$ where τ is known and $\mu\epsilon[\mu_1, \mu_2]$. (Observe that this reduces to the well-known Bayesian case when $\mu_1 = \mu_2$).

A trite calculation shows that the extremal posterior set of distributions is the subset of normal distributions $N(\mu',\tau')$, where

$$\mu' \epsilon \left[\frac{\tau\mu_1 + nr\bar{x}}{\tau + nr} ; \frac{\tau\mu_2 + nr\bar{x}}{\tau + nr} \right]$$

and

$$\tau' = \tau + nr \quad \text{with } \bar{x} = \frac{\Sigma x_i}{n}$$

Then the quasi-Bayesian confidence interval for w for a given confidence coefficient p is

$$\left[\frac{\tau\mu_1 + nr\bar{x}}{\tau + nr} - \lambda_p \left(\frac{1}{\tau + nr}\right)^{1/2} ; \frac{\tau\mu_2 + nr\bar{x}}{\tau + nr} + \lambda_p \left(\frac{1}{\tau + nr}\right)^{1/2} \right],$$

where

$$\Phi(\lambda_p) - \Phi(-\lambda_p) = p.$$

Observe that any of the distributions of the posterior set (not only the extremal ones which are normal) assigns to this interval a probability greater than p.

Let us now see how this interval compares with a Bayesian confidence interval for any prior distribution compatible with our partial information.

Suppose the prior distribution is $N(\mu,\tau)$ with $\mu\epsilon[\mu_1, \mu_2]$. For a sample size n' the Bayesian p-confidence interval is

$$\left[\frac{\tau\mu + n'r'\bar{x}'}{\tau + n'r} - \lambda_p \left(\frac{1}{\tau + nr'}\right)^{1/2} ; \frac{\tau\mu + n'r\bar{x}'}{\tau + n'r} + \lambda_p \left(\frac{1}{\tau + n'r}\right)^{1/2} \right],$$

where $\bar{x}' = \dfrac{\Sigma x_i}{n'}$

It is evident that for the same sample-size the quasi-Bayesian interval is wider than the corresponding Bayesian one. If we now equate the width of the two intervals for sample sizes n and n' respectively we obtain

$$\frac{\tau(\mu_2-\mu_1)-2\lambda_p}{\tau + nr} = \frac{2\lambda_p}{\tau + n'r}$$

It is interesting to note that the above relation does not depend on $\bar{x}, \bar{x}'$ and it obviously implies that $n \geq n'$. The difference n-n' could be interpreted as the "additional sample size" for which partial prior information could be considered as total prior information.

b) Quasi-Bayesian estimators for the mean of a normal distribution

Suppose the same situation of normal sampling as in example a) with the same partial information. If the loss function for this decision problem is

$$L(w, d) = (w - d)^2$$

the quasi-Bayesian estimator is seen to be

$$\delta^*(x_1,\ldots,x_n) = \left[\frac{\tau\mu_1 + nr\bar{x}}{\tau + nr} ; \frac{\tau\mu_2 + nr\bar{x}}{\tau + nr} \right] \qquad (4.1)$$

which reduces to a single point if either μ_2-$\mu_1 \to 0$, $\tau \to 0$, or $n \to \infty$

It deserves mentioning that the quasi-Bayesian estimator in this case is the union of Bayes estimators corresponding to the extremal posterior distribution. Any Bayesian estimator corresponding to a non extremal posterior distribution belongs to δ^*.

Note that if partial information reduces to the following: "Prior information is normal $N(\mu, \tau)$ with $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$, $0 \leq \lambda \leq 1$", the quasi-Bayesian estimator is the same as the one given by (4.1)

c) Quasi-Bayesian testing of hypotheses

In this section we consider the simplest example of testing a simple null hypotheses versus a simple alternative hypotheses, so that the two states, two actions, decision problem is,

	θ_0	θ_1
a_0	o	a
a_1	b	o

$a,b>0$

where θ_0 stands for the null hypotheses and θ_1 for the alternative; a_0 accept θ_0 and a_1 reject θ_0 (and accept θ_1, accordingly)

Partial information in this example is given in the form of a closed interval that represents the range of possible values of prior probability on the null hypotheses, that is

$$P\,[\theta_0] \in [\xi_0, \xi_1] \qquad\qquad (0 \leq \xi_0 \leq \xi_1 \leq 1)$$

If we represent the density (with respect to some dominating measure) of a sample of size one, when θ_i $(i=0,1)$ is true by f_i, then the quasi-Bayes procedure for this decision problem when a random sample of size n is taken, which we could name "quasi-Bayesian test", is the following

$$\delta^*(x_1,\ldots,x_n) = \begin{cases} a_0 \text{ if } \Pi_{i=1}^{n} \dfrac{f_1(x_i)}{f_0(x_i)} \leq \dfrac{b}{a}\,\dfrac{\xi_0}{1\text{-}\xi_0} \\[2ex] a_1 \text{ if } \Pi_{i=1}^{n} \dfrac{f_1(x_i)}{f_0(x_i)} \geq \dfrac{b}{a}\,\dfrac{\xi_1}{1\text{-}\xi_1} \\[2ex] \{a_0, a_1,\} \text{ if } \dfrac{b}{a}\,\dfrac{\xi_0}{1\text{-}\xi_0} < \dfrac{\Pi_{i=1}^{n} f_1(x_i)}{\Pi_{i=1}^{n} f_0(x_i)} < \dfrac{b}{a}\,\dfrac{\xi_1}{1\text{-}\xi_1} \end{cases}$$

This results needs some explanation: If the sample observed is such that $\delta^*(x_1,\ldots,x_n)$ equals a_0 or a_1, there is no problem, and the null hypotheses is accepted or rejected, respectively. If, however, $\delta^*(x_1,\ldots,x_n) = \{a_0, a_1\}$ then no single course of action is possible.

This means that our partial (posterior) information is not enough as to discriminate between the two actions so that new sample information is needed and a computation of the new likelihood ratio may show that $\delta^*(x_1,\ldots,x_n, x_{n+1})$ equals either a_0 or a_1 or if $\delta^*(x_1,\ldots,x_n,x_{n+1}) = \{a_0, a_1\}$ a new sample is required, and so on. This brings out the strong analogy between the quasi-bayesian test and Wald's sequential probability ratio test with barriers

$$A = \frac{b}{a}\,\frac{\xi_0}{1\text{-}\xi_0} \qquad \text{and } B = \frac{b}{a}\,\frac{\xi_1}{1\text{-}\xi_1},$$

in the case the cost of new observations is not included within the structure of the decision problem.

REFERENCES

AUMANN, R.J. (1962). Utility theory without the completeness axiom. *Econometrica* **30**, 445-462.

— (1964). Utility theory without the completeness axiom: a correction. *Econometrica* **32**, 210-212.

BLACKWELL, D. and GIRSHICK, M.A. (1954). *Theory of Games and Statistical Decisions.* Wiley: New York.

CRIADO, F. (1978). *Algunas caracterizaciones de la utilidad y extensiones.* Ph.D. Thesis. Universidad de Málaga.

DEMPSTER, A.D. (1967). Upper and lower probabilitites induced by a multivalued mapping. *Ann. Math. Statist.* **38**, 325-339.

— (1968). A generalization of Bayesian Inference. (with discussion). *J. Roy. Statist. Soc. B* **30**, 205-232.

FINE, T. (1973). *Theory of Probability: An Examination of Foundations.* Academic Press: New York.

FISHBURN, P.C. (1964). *Decision and Value Theory.* Wiley: New York.

— (1975). Weak comparative probability on infinite sets. *Ann. Prob.* **3**, 889-893.

GIRON, F.J. (1975). S-juegos generalizados. *Rev. Real Acad. Ciencias Madrid* **69**, 49-97.

— (1977). Caracterización axiomática de la regla de Bayes y la probabilidad subjetiva, *Rev. Acad. Cienc. Madrid.* **71**, 19-101.

— (1978). Una caracterización de la incertidumbre parcial. *Actas V Jornadas Luso-Espanholas de Matemática.* Aveiro, Portugal.

— (1979). Probabilidad y utilidad: conceptos duales de la teoria de la decisión. *Rev. Real Acad. Cienc. Madrid,* **73**, 225-230.

GOOD, I.J. (1962). The measure of a non-measurable set. In *Logic, Methodology and Philosophy of Science.* E. Nagel, P. Suppers and A. Tarski eds. 319-329. Standford: University Press.

GOODMAN, T.N.T. (1977). Qualitative probability and improper distributions. *J. Roy. Statist. Soc. B* **39**, 387-393.

KOOPMAN, B.O. (1940). The axioms and algebra of intuitive probabilty. *Ann. Math.* **41**, 269-292.

LEITMANN, G. (1976). *Multicriteria Decision Making and Differential Games.* London: Plenum Press.

RIOS, S. (1975a). Nuevos criterios de ordenación de reglas de decisión. *Trab. Estadist.* **26**, 5-12.

— (1975b). Ordre quasi-bayesien des regles de decision. *Proc. 40th Session of I.S.I. Warsaw,* 694-698.

— (1976). Nuevos criterios de ordenación de reglas de decisión. *Rev. Real Acad. Cienc. Madrid* **70**, 235-253.

SAVAGE, L.J. (1954). *The Foundations of Statistics.* Wiley: New York.

SMITH, C.A.B. (1961). Consistency in statistical inference and decision (with discussion). *J. Roy. Statist. Soc. B* **23**, 1-25.

— (1965). Personal probability and statistical analysis (with discussion). *J. Roy. Statist. Soc.* **A 128**, 469-499.

SUPPES, P. (1974). The measurement of belief (with discussion). *J. Roy. Statist. Soc. B* **36**, 160-175.

On Some Statistical Paradoxes and Non-conglomerability

BRUCE M. HILL
University of Michigan
and
University of Utah

SUMMARY

Some statistical paradoxes arising from the use of non-conglomerable finitely additive distributions are discussed.

Keywords: FINITE ADDITIVITY; CONGLOMERABILITY; LIKELIHOOD PRINCIPLE.

1. INTRODUCTION

In recent years there has been a revival of interest both in statistical paradoxes, and in the finitely additive theory of Bruno de Finetti(1972, 1974), Dubins (1975), Heath and Sudderth (1976), Hill (1975), Lane and Sudderth (1978), Stone (1976). Most such paradoxes are transparent from the finitely additive subjective Bayesian point of view, and require little comment. The few that remain are essentially paradoxes of non-conglomerability, a concept due to de Finetti (1972, p. 204), (1974, p. 177). These appear to be real paradoxes and stretch the imagination. At present the finitely additive theory of de Finetti is the only theory of inference and decision-making without gaping holes, and it therefore is important to clarify such paradoxes so far as possible.

In this article I first discuss Mervyn Stone's example (Stone, 1976) of what he calls "strong inconsistency from uniform priors", from the finitely additive subjective Bayesian point of view. The example is both interesting and important. Its importance, of course, does not depend upon the extent to which it arises in real life, but rather upon what it tells us about modes of inference and decision-making. If, for example, the trap that Stone has set for the unwary Bayesian had real teeth, then one might be forced to regard both improper and finitely additive prior distributions as potentially dangerous, and either to abandon the Bayesian approach entirely, or at least to restrict its use to very special situations. Even the latter would sacrifice one of the great advantages of the Bayesian approach, namely, its universality, as opposed to other theories of inference (fiducial, Neyman-Pearson), which break down in all but the simplest problems. Unfortunately both Stone's presentation and the discussion seem to have obscured the real issues. I would like therefore to present the example in my own notation, and raise and discuss the issues from the finitely additive point of view. This will lead us to non-conglomerability, and perhaps a real paradox.

2. LADY AND DRUNKEN SAILOR

The setting is flatland, laid out in blocks as in Stone (1976). Starting from a known origin a lady and a drunken sailor walk about, trailing a string behind them. The path traced by their string consists of vertical and horizontal line segments, and whenever a block is retraced the string is pulled tight, so that such retracings are not visible. Eventually they stop at an intersection and bury a treasure. A mechanism is then used to select a direction, with each assumed equally likely, and the pair walk one block in the chosen direction, still trailing the string. The sailor then dies on the spot and is buried there by the lady, who disappears into the night. No other information is provided as to the manner in which the lady and drunken sailor have ambled, and the data of the experiment consists of the tight path from the origin to the sailor's grave.

Let $\hat{x}$ be the point at which the sailor is buried, and let $\hat{p}$ be the tight path from the origin to $\hat{x}$, so that $\hat{p}$ is the data, and $\hat{x}$ is the endpoint of $\hat{p}$. Let X be the true location of the treasure, and let P be the true path from the origin to X. Now let $\hat{x}_1$ be the point one block back from $\hat{x}$ along $\hat{p}$, and let $\hat{p}_1$ be the observed path from the origin to $\hat{x}_1$.
Let $\hat{x}_2$, $\hat{x}_3$, $\hat{x}_4$, be taken counterclockwise around $\hat{x}$, starting from $\hat{x}_1$, so that the $\hat{x}_i$, $i = 1, 2, 3, 4$, are the four points surrounding $\hat{x}$. Finally, let $\hat{p}_i$ be the tight path extending $\hat{p}_1$ so as to pass through $\hat{x}$, and then through $\hat{x}_i$, $i = 2, 3, 4$. Thus, such $\hat{p}_i$ are exactly two blocks longer than $\hat{p}_1$, and given the data $\hat{p}$, it is known with certainty that the true path to the treasure is one of the $\hat{p}_i$, $i = 1, 2, 3, 4$, and that the location of the treasure is one of the $\hat{x}_i$, $i = 1, 2, 3, 4$.

Let us determine the posterior odds for $\hat{x}_1$ versus $\hat{x}_i$, $i = 2, 3, 4$ using the finitely additive approach. We condition upon the data $\hat{p}$. Although such a conditioning event may have subjective probability 0, the conditional probabilities are still well defined in the de Finetti theory (de Finetti, 1972, p.82, 1974 p.173). Since, given $\hat{p}$, the event $X = \hat{x}_i$ is equivalent to the event $P = \hat{p}_i$, then provided the ratio is not indeterminate,

$$\frac{Pr\{X = \hat{x}_1|\hat{p}\}}{Pr\{X = \hat{x}_i|\hat{p}\}} = \frac{Pr\{P = \hat{p}_1|\hat{p}\}}{Pr\{P = \hat{p}_i|\hat{p}\}} \tag{2.1}$$

$$= \frac{Pr\{P = \hat{p}_1|\hat{p}_1,\hat{p}_2,\hat{p}_3,\hat{p}_4\}}{Pr\{P = \hat{p}_i|\hat{p}_1,\hat{p}_2,\hat{p}_3,\hat{p}_4\}}, \qquad i = 2,3,4.[1]$$

It is essential to note that $Pr\{ P = \hat{p}_i\}$ is simply the unconditional *a priori* subjective probability that the true path to the treasure is the particular path $\hat{p}_i$. Thus the Bayesian solution depends entirely upon the specification of the prior distribution for P, and no solution can be obtained without such a specification, whether explicit or implicit. Before examining some interesting specifications of the prior distribution for P let us stop and see how use of the parameter X instead of P led Fraser to some amusingly tragic (or tragically amusing) conclusions.

In his discussion Fraser felt that Stone had missed the real point of the example, namely that it was perhaps devestating evidence against the likelihood principle itself. Let us try to see how Fraser might have been led to such a conclusion. Suppose that instead of the data $\hat{p}$ the data had consisted only of the directed line segment, say $\vec{x}$, from $\hat{x}_1$ to $\hat{x}$. Call this experiment $\vec{E}$. Then the likelihood function derived from $\vec{E}$, with data $\vec{x}$, would be $\vec{L}(\hat{x}_i^{\rightarrow}) = ¼$, for $i = 1, 2, 3, 4$, and would be 0 elsewhere. Here $\hat{x}_1^{\rightarrow} = \hat{x}_1$, but for $i = 2, 3, 4$, we have replaced the treasure location $\hat{x}_i$ by the corresponding last two *segments of $\hat{p}_i$, i.e., the directed segment from $\hat{x}_1$ to $\hat{x}$ followed by the directed* segment from $\hat{x}$ to $\hat{x}_i$. Given $\vec{x}$, specification of the treasure location parameter X is equivalent to specification of the last two line segments of P. Now consider still a third experiment $\dot{E}$ in which things are as before except that only the sailor's burial point, $\hat{x}$, is available. The likelihood function for X derived from $\dot{E}$ is then $\dot{L}(\hat{x}_i) = ¼$, $i = 1, 2, 3, 4$. In arguing against the likelihood principle Fraser apparently views the likelihood function $\vec{L}(\hat{x}_i^{\rightarrow})$ derived from $\vec{E}$ as identical with $\dot{L}(\hat{x}_i)$ derived from $\dot{E}$. Strictly speaking this is not valid, since $\vec{L}(\bullet)$ and $\dot{L}(\bullet)$ are defined on different spaces. Thus Fraser's argument would apply, at best, only to a generalized version of the likelihood

[1] This evaluation is based upon de Finetti's Axiom 3 [1974, Vol. 2, p. 399], which asserts that conditional probabilities satisfy the first two axioms. One can then evaluate ratios of conditional probabilities even when the conditioning event has probability zero. Thus (2.1), which is trivial when $\hat{p}$ has positive probability, can still be obtained, by repeated use of Axiom 3, even when $\hat{p}$ has probability zero. The method is to condition events like $P = \hat{p}_i$ upon $\hat{p}$ and ($\hat{p}_1$ or $\hat{p}_i$), so that 0/0 cannot occur.

principle. But this generalized version would be unacceptable to Bayesians also since in fact the posterior distribution of X derived from $\vec{E}$ is in general different from that derived from $\dot{E}$. Note that when the prior distribution for X is uniform then $\dot{E}$ gives rise to a uniform posterior distribution over the $\hat{x}_i$, i.e., to what I shall call the Stoned Bayesian Posterior. There is an additional flaw in Fraser's argument against the likelihood principle, namely, that the original experiment is not equivalent to $\vec{E}$ unless $\hat{p}_1$ contains no information about X. This, however, need not be the case, and furthermore $Pr\{\hat{p}_1 | X = \hat{x}_i\}$ is not even well defined without at least a partial specification for the prior distribution of P. As we shall see later, prior distributions for which $Pr\{\hat{p}_1 | X = \hat{x}_i\}$ is constant, have some peculiar features. From a more general point of view note that the likelihood function $L(p) = Pr\{\hat{p} | P = p\}$, in the original experiment, is a function whose domain consists of all possible paths to the treasure, while $\dot{L}(x)$ is a function whose domain consists of all possible treasure burial points. Although, given $\hat{p}$, $P = \hat{p}_i$ if and only if $X = \hat{x}_i$, $i = 1, 2, 3, 4$, the likelihood functions cannot be identified in the two experiments.

Now let us try to formulate a prior distribution for P. The model I find most compelling is as follows. Suppose that very little can be presumed as to the walking rate of the lady and sailor. They may, for example, at times stop somewhere discretely, and at other times may run. By symmetry we might view all tight paths of a given length (number of tight string segments, or blocks walked, exclusive of retracings) as equally likely apriori. Let N be the true length of the tight path to the treasure, and let $\hat{n}$ be the observed length of the tight path to $\hat{x}$. Then we need only specify a prior probability distribution $q(j) = Pr\{N = j\}$, $j = 0, 1, 2, \ldots$ According to the de Finetti theory any finitely additive distribution can be used, including, of course, countably additive ones. Supposing it is known that the time during which the lady and sailor amble is not so small as to be very informative, and that similarly $\hat{n}$ is not too small, we might wish to take $Pr\{N = \hat{n} - 1\} \approx Pr\{N = \hat{n} + 1\}$. Who would, for example, wish to regard paths of length 1000 blocks as much more or less probable than paths of length 1002 blocks, over a not insubstantial length of time? Thus we shall assume $q(\hat{n} - 1) \approx q(\hat{n} + 1)$ for the given $\hat{n}$. But for any specified path of length $j - 1$, there are exactly 9 tight paths of length $j + 1$, which continue the given path by two blocks. By symmetry this means that any particular such path of length $j + 1$ must have one ninth the apriori probability of the specified path of length $j - 1$. From (2.1) it now follows that

$$\frac{Pr\{X = \hat{x}_1 | \hat{p}\}}{Pr\{X = \hat{x}_i | \hat{p}\}} = \frac{Pr\{P = \hat{p}_1 | \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4\}}{Pr\{P = \hat{p}_i | \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4\}} = \frac{9\,q(\hat{n} - 1)}{q(\hat{n} + 1)}$$

$$\approx 9, i = 2, 3, 4.$$

This, of course, yields $Pr\{X = \hat{x}_1 | \hat{p}\} \approx 3/4$, and corresponds more or less to the confidence solution proposed by Stone. Apart from our use of changing walking rates, and the time factor, both of which suggest taking $q(\hat{n} - 1) \approx q(\hat{n} + 1)$, this analysis is similar to that proposed by Dickey in his discussion of the Stone article.

Although the evaluation $q(\hat{n} - 1) \approx q(\hat{n} + 1)$ seems compelling, as mentioned above the de Finetti theory allows use of any finitely additive distribution for N. To obtain the Stoned Bayesian Posterior under our assumption of symmetry, one would thus need $q(\hat{n} + 1) = 9\,q(\hat{n} - 1)$. Contrary to Stones's analysis, it is not a uniform prior distribution on X, nor even a uniform prior distribution on N, that is relevant to Stone's trap, but rather the distribution for which $q(j + 1)/q(j - 1) \equiv 9$, $j = 2, 3, \ldots$ There do exist finitely additive distributions with this property. For example, let J have the usual uniform finitely additive distribution over the non-negative integers, i.e., an integer "chosen at random" in the de Finetti terminology de Finetti, (1972, p.86). Let

$\tilde{0} = \{0\}$, $\tilde{1} = \{1, 2, 3\}$, $\tilde{2} = \{4, 5, 6, 7, 8, 9, 10, 11, 12\}$ etc. If $\tilde{q}(j) = Pr\{J \in \tilde{j}\}$, then using de Finetti's Axiom 3 (1974, Vol. 2, p. 399), or results of Dubins (1975), we can define the conditional ratio as

$$\frac{\tilde{q}(j + 1)}{\tilde{q}(j - 1)} \qquad = 9, j = 2, 3, \ldots^{2}$$

Thus if there were a serious argument against the use of $Pr\{N = j\} = \tilde{q}(j)$ as a prior distribution, then this would speak against the finitely additive theory. And so we should squarely face up to the question as to whether use of $\tilde{q}(\cdot)$ (no matter how unnatural and uncompelling) leads to any unfortunate consequences. Clearly use of $\tilde{q}(\cdot)$ leads to a coherent procedure in the sense of de Finetti, so there is no possibility of being made a sure loser.

Let us now look at the nature of Stone's argument against the Stoned Bayesian. We consider two individuals, S and $S.B.$, each of whom can search in exactly three places for the treasure. S always chooses $\hat{x}_1, \hat{x}_2, \hat{x}_3$, while $S.B.$ always chooses $\hat{x}_2, \hat{x}_3, \hat{x}_4$. In repeated experiments Stone suggests that S can be "confident" of obtaining the treasure in at least about 3/4 of the experiments, while $S.B.$ can be so "confident" in at most about ¼ of the experiments.

[2] This again follows from Axiom 3, with

$$\frac{Pr\{j + 1\}}{Pr\{j - 1\}} \text{ defined as } \frac{Pr\{j + 1 \mid j - 1 \text{ or } j + 1\}}{Pr\{j - 1 \mid j - 1 \text{ or } j + 1\}}$$

Apparently none of the discussants questioned this "confidence". I would like to do so. The basis of Stone's argument is, of course, the fact that conditional upon $X = x$, there is probability 3/4 that the mechanism for choosing a direction will choose one that extends the path. Since this is true for all possible x, Stone apparently draws the conclusion that 3/4 is also appropriate unconditionally. A serious discussion of this question necessarily leads us to the concept of nonconglomerability.

Let us begin by considering a different example, also paradoxical. A point is selected uniformly on the surface of a sphere with a designated north and south pole. You are given the **exact** longitude of the point relative to some specified great circle, and can choose to search for the point either in the equatorial arc between 45° north and south latitude, or alternatively along the corresponding polar arcs. Assume that you are certain to find the point if you search along the arc in which it lies, and suppose you always choose the polar arcs for your search. There are serious arguments for regarding the point as uniform over the possible points compatible with the given longitude[3] (de Finetti, 1974, p. 275) (as de Finetti argues, the Kolmogorov resolution of the paradox in terms of limiting surface areas is merely an artifice to avoid the logical issues). In this case, given the longitude, say $X = x$, the probability that you find the point is ½, for each possible x. Stone's frequency argument would then lead you to anticipate that in approximately ½ of such experiments you will find the point. On the other hand, in terms of surface areas (not conditional on longitude), one might argue that your frequency of finding the point should be substantially less than ½, namely, $1 - 1/\sqrt{2}$.

Which frequency, if either, is relevant? Note that we are not here raising questions as to the subjective versus frequency concepts of probability, but for the present accepting the frequency interpretation, and arguing that even within its own framework it does not yield an unambiguous anticipated frequency of success. By the same token neither does Stone's argument give an unambiguous result for the frequency of finding the treasure. (Some may try to avoid the dilemma in the case of the sphere via the Kolmogorov fashion (Kolmogorov, 1950, p. 50), or by arguing that "real" problems are discrete or even finite. Such arguments merely avoid the logical content of the problem. Furthermore, even if we accept that real problems are finite, continuous idealizations are commonly made in statistics for practical approximations, so the question would remain as to when such idealizations are dangerous).

de Finetti long ago described such situations in terms of non-conglomerability and argued that the paradoxes are real rather than apparent. In the finitely additive theory it is possible that

$$Pr\{\text{Find treasure} \mid X = x\} = C, \text{ for all possible } x, \text{ but } Pr\{\text{Find treasure}\} \neq C.$$

My point is that there is a very weak link in Stone's argument, which is equivalent to an assumption of conglomerability for relative frequencies in repeated experimentation. The sphere example makes it clear that a real assumption is involved in such reasoning. Note, incidentally, that if Stone had taken a finite flatland (say walled around), then his argument against Stoned Bayesian would not apply, even apart from non-conglomerability considerations, so that the infinite idealization is essential to the example.

3. A REAL PARADOX?

Suppose we are concerned with the value of a physical constant M, known to be rational and between 0 and 1. Suppose further that two different physical theories are under consideration, T_1 and T_2. Given T_1, M has some specified distribution Q_1, concentrated on the rationals in the unit interval, and giving each such rational positive probability. Given T_2, M has the diffuse finitely additive uniform distribution on the rationals between 0 and 1, and thus gives each such rational zero probability. We assume $0 < Pr\{T_1\} < 1$, $Pr\{T_2\} = 1 - Pr\{T_1\}$. You are now given the exact value of M, say $M = m$, and wish to reassess the probabilities of the two theories on the basis of this data. Note that $Pr\{M = m\} = Pr\{T_1\} Pr\{M = m \mid T_1\} > 0$, so the data upon which we condition has positive prior probability. The de Finetti theory then yields

$Pr\{T_1 \mid M = m\} = 1$, for each possible m. This is another example of non-conglomerability, since by assumption $Pr\{T_1\} < 1$. However, there is a further paradoxical aspect. For one knows in advance that no matter what rational

[3] Of course the Kolmogorov axioms rule out such conditional uniformity on great circles, so it is a question of an appropriate choice of axioms. Consider, however, the following coordinate free formulation. Suppose the sphere is not labelled with a prespecified north pole, and you regard the point as uniformly distributed on the surface of the sphere. If you are given *only* the information that the point lies on some *exact* great circle, would you now regard the point as uniform on that great circle (since there is no north pole, no other distribution seems natural, so presumably you either regard the distribution as uniform, or else consider it as indeterminate)? Is this case necessarily different from that in which there is a prespecified north pole? If your answer is yes, try the following variation. Before observing the data (i.e., the great circle on which the point lies) *every* great circle on the sphere is labelled with a north pole by means of the axiom of choice. When you are given the data the sphere is rigidly rotated in a prescribed fashion, so that the chosen north pole for this particular circle is in some standard position, i.e., agrees with a standard north pole and is at the longitude of Greenwich. The (thought) experiment is then repeated n times, independently, and so one obtains in this way n points on the great circle through Greenwich. How do you view these points as being distributed on the great circle through Greenwich? Does latitude have the conventional cosine density, or are the points distributed uniformly, or is no opinion justified? Is this case necessarily different from that in which there is a single prespecified north pole?

value m is observed, it is a foregone conclusion that T_1 will have posterior probability 1. This example is similar in character to that of Dubins (1975), discussed by de Finetti (1972, p. 205). (Other examples of de Finetti, for instance, the probability that an integer is even, conditional on any element of a partition (de Finetti, 1974, p. 178), are less extreme, since although this probability may be 1 for each such element, there also exist partitions for which it is identically 0. Thus the conclusion is only foregone with respect to a specified partition).

A number of remarks seem pertinent.

1. The problem we have posed arises frequently in statistical practice. In the conventional Bayesian formulation it would be a matter of comparing prior distribution Q_1 for a Bernoulli parameter p, as against a uniform prior on p. Note that if Q_1 remains concentrated on the rationals, but under $T_2, p \sim U[0, 1]$, then the paradox disappears, since if m is irrational the posterior probability of T_2 becomes 1. Imagine that initially the problem was posed with $M \sim U[0, 1]$, given T_2, but that before observing m it was learned that irrational values of M are impossible. Our initial description of the problem in terms of a physical constant rather than a Bernoulli p was chosen so as to avoid issues regarding the subjective interpretation of such a parameter as p.

2. If we take the Bernoulli p version of the problem, but replace the original form of data by a finite number of observations on a Bernoulli sequence with parameter p, then the paradox again disappears. The posterior distribution of future observations can be described in the usual way as a nondegenerate mixture under T_1 and T_2.

3. The problem is related to that formulated by Harold Jeffreys (1967, Ch. V) with regard to testing $u = 0$ versus u having the Jeffreys uniform improper prior density, where u is the mean of a normal population. Given normally distributed observational data on u, one would ordinarily reject the hypothesis that $u = 0$. This led Jeffreys to regard use of the improper prior as inappropriate for "hypothesis testing" purposes, although he retained it for estimation. Jeffreys's model for hypothesis testing was developed by L.J. Savage in terms of testing a sharp null hypothesis against a diffuse alternative, and later Hill (1975) formalized these notions into a unified structure involving various hypotheses, each given positive prior probability, and conditional upon each of which, certain parameters are given proper prior distributions. The Jeffreys paradox is much the same as that concerning M, and Jeffreys, Savage and Hill avoided the difficulty by taking a proper prior distribution for the parameters under the alternative hypothesis. From a practical point of view this may be satisfactory, but there remain logical questions as to the use of a finitely additive prior under the alternative.

Now let us return to the problem as initially formulated. What are the implications of the paradox? A first reaction might be as follows. Since any observation m will lead one to attach posterior probability 1 to T_1, it is unnecessary to make such an observation, and one can simply alter $Pr\{T_1\}$ to be 1 even without performing the experiment. (Note that this argument would not apply in the other de Finetti examples alluded to above, since the conclusion would depend upon the chosen partition). In effect this line of argument repudiates the diffuse finitely additive distribution on the rationals, at least for "testing" purposes, much as Jeffreys repudiated the uniform improper prior on u. The argument against such repudiation is as follows. It is certainly possible that a physical theory, such as T_2, can lead one to regard M as uniform over the rationals, and under the subjective interpretation the prior probability for T_2 can then be any value between 0 and 1. Suppose that you evaluate $Pr\{T_2\} > 0$. Is it possible that merely by contemplating the experiment which consists of observing M, that one can conclude that one should have evaluated $Pr\{T_2\} = 0$? This would appear disastrous for the subjectivistic theory, since it would imply that the prior probability of a hypothesis could not be assessed unless one knew beforehand what experiment would be performed. It would also suggest that if some other experiment was later even contemplated, then the prior probability of the hypothesis might have to be revised. Surely any subjectivistic analysis would be impractical in this case. Thus if we wish to retain the subjectivistic theory, including finitely additive diffuse models, then we must learn to live with nonconglomerability. In particular, we must accept that we can have $Pr\{T_2 | M = m\} = 0$ for all possible m, while $Pr\{T_2\} > 0$. This is only a stronger form of the paradox of the sphere example. In that example, given any longitude, one regards the polar arcs as having conditional probability ½. The same argument that suggested taking $Pr\{T_2\} = 0$ because $Pr\{T_2 | M = m\} = 0$ for all m, would suggest in the sphere example that the union of the polar arcs, i.e., the polar caps, should have prior probability ½, contrary to the initial assumption that the point was chosen uniformly on the surface of the sphere. Yet surely no one would use this as an argument to repudiate the uniform distribution on the sphere.

4. FINAL COMMENTS

1. It is desirable that the subjectivistic theory should be capable of dealing with models in which the parameter can take on infinitely many values, with cardinality irrelevant. This is so in part for purely logical reasons, so that the theory forms a coherent structure that can be relied upon in all situations; and in part for practical reasons, since even from the ordinarily more realistic finitistic point of view, one will often find it advisable to make

approximations using infinite models.

2. Just as in the finite case, so too in the infinite case, there are sometimes compelling psychological reasons for choosing prior distributions that are, in some appropriate sense, "uniform" over the possible parameter values. Of course the choice of parameter for which uniformity seems appropriate (whether exact or approximate) is often subtle. Thus in Stone's example uniformity on the "length" of path seems psychologically natural, just as in the context of multivariate inadmissibility uniformity on the norm of the parameter vector seems natural (Hill, 1975). Within the subjectivistic theory such uniformity is neither mandatory nor excluded.

3. The conventional use of improper prior densities to represent uniformity (or so-called "ignorance") can lead to difficulties with regard to coherence and admissibility (Heath & Sudderth, 1976), (Hill, 1975). On the other hand, the use of finitely additive prior distributions, as advocated by de Finetti, is part of a solid theoretical framework for Bayesian inference and decision-theory. If one uses posterior distributions for finitely additive prior distributions then one cannot be made a sure loser, nor even to have uniformly positive expected loss, in a finite number of bets regarding parameters, and the corresponding decision procedures are extended admissible for bounded loss functions (Heath & Sudderth, 1976). The only difficulties known to this author with regard to the use of finitely additive prior distributions are those that arise from non-conglomerability, as in the example of Stone, that of Section 3, and that of uniformity on the surface of the sphere. Our discussion of these examples has been an attempt to show not only that non-conglomerability is unavoidable, but also that even within the frequentist theory, frequency arguments such as that of Stone may be unconvincing precisely because of non-conglomerability.

For those of us who wish to retain the subjective Bayesian model for learning and decision-making there appear to be three main paths open. First, we can restrict the model to finitistic applications and/or to bounded loss functions and proper prior distributions footnote in the infinite case; second, we can persist with conventional improper prior distributions[4] in the infinite case, ignoring inadmissibility (and even extended inadmissibility) problems; third, we can develop the finitely additive theory, learning to live with non-conglomerability. The first path is quite restrictive and may be unrealistic even as an approximation. The second path, at least in some applications, will lead to unnecessary losses. Are there any real objections to the third path?

[4] Much of Jeffreys' use of improper prior distributions can in fact be justified by the finitely additive theory using de Finetti's Axiom 3. I find the de Finetti approach preferable in that the axioms have a very clear intuitive content. Thus Axioms 1 and 2 merely formalize the aim of not being a sure loser, while Axiom 3 articulates this aim with the events of zero probability. It is not clear to me what are the corresponding aims of the countably additive improper prior approach of Jeffreys. For example, how should one view inadmissibility and the lack of extended admissibility in Jeffreys' approach? The de Finetti approach clearly allows inadmissibility, and if one chooses non-conglomerable distributions, it even allows the lack of extended admissibility. It thus provides a clear framework for discussion.

REFERENCES

DE FINNETI, B. (1972). *Probability, Induction and Statistics*, New York: Wiley.

— (1974). *Theory of Probability, Vol.* 1 and 2. New York: Wiley.

DUBINS, L.E. (1975). Finitely additive conditional probabilities, conglomerability, and desintegrations. *Ann Probability* 3, 89-99.

HEATH, D. and SUDDERTH, W.D. (1976). On finitely additive priors, coherence, and extended admissibility. *Ann. Statist.* 6, 333-345.

HILL, B.M. (1975). On coherence, inadmissibility, and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, (S.E. Fienberg and A. Zellner, eds.), Amsterdam: North-Holland, 555-584.

LANE, D. and SUDDERTH, W.D. (1978). Diffuse models for sampling and predictive inference. *Ann. Statist.* 6, 1318-1336.

JEFFREYS, H. (1967). *Theory of Probability*, Oxford: University Press.

KOLMOGOROV, A.N. (1950). *Foundations of the Theory of Probability*. New York: Chelsea.

STONE, M. (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.* 71, 114-116.

DISCUSSION

I.J. GOOD *(Virginia Polythecnic and State University):*

I shall restrict my discussion to the historical aspects of the paper presented by Professors Giron and Rios.

J.M. Keynes (1921) argued that not all logical probabilities could be compared. B.O. Koopman (1940a, b), acknowledged Keynes's influence and laid down fairly convincing but complicated axioms for partially ordered "intuitive" probabilities, where "intuitive" I think meant either logical or subjective. I propounded the simplest possible acceptable theory of partially ordered subjective probabilities in Good (1950, p. 119) and pointed out that such a theory is identical with the use of upper and lower probabilities provided that it is agreed that we can imagine perfectly shuffled packs of cards. I extended the theory to include utilities in an obvious manner in Good (1952) or see Good (1954). At a 1960 conference in Stanford (Good, 1962) I showed that this simplest possible theory of partially ordered probabilities implies formal axioms connecting upper and lower probabilities. Cedric Smith (1961) justified my theory by using arguments analogous to those used by Savage (1954) for the theory of sharp probabilities. For this justification he made use of convex sets of prior distributions. Smith said he left some loose ends and presumably these have received the attention of Rios and Giron. Whether this is so I have unfortunately not had time to check.

Most of my historical research was concerned with finding the publications where I had mentioned the partially ordered theory of subjective probability. I have found fifty such publications, or perhaps only 49, ranging from 1949 to 1979. I have given the list to the authors, to prove to them that I have emphasized the partially ordered theory perhaps *ad nauseam*, but in the references to this discussion I have listed only Good (1950, 1952, 1962, 1976 and 1977). For example, in Good (1976, p. 137) I pointed out that my theory is a Bayes/non Bayes compromise, as Rios and Giron have now recognized.

It may be helpful to mention that the theory of partially-ordered probabilities (and utilities) is sometimes called a theory of comparative or qualitative probabilities (and utilities). The subjective version could reasonably be called Good's theory or the Doogian theory or the comparative or qualitative or partially-ordered Bayesian theory, and "quasi-Bayesian" is yet another name for the same thing.

Although I have always accepted this theory, in practice I often prefer to use sharp probabilities and utilities for the sake of simplicity, as an approximation to the partially ordered theory.

On a point of terminology, I think the expression "confidence interval" should be restricted to the Neyman-Pearson sense. In the Bayesian theory one can use the expression "Bayesian estimation interval".

Turning now to Professor Hills' paper, the word "paradox" has at least two distinct meanings which can be distinguished by talking about apparent and true paradoxes. If I thought there were any true paradoxes in the theory of subjective probabilities that I support, then I would be forced to abandon rationality. I am not yet prepared to do that.

Perhaps the common denominator of all Bayesian statistics is the product law, $P(A \& B) = P(A).P(B|A)$, meaning that if two of the probabilities mean anything then so does the third, and this is so even if the probabilities are merely constrained by inequalities. Have we any reason to doubt this product law, in the light of the various apparent paradoxes mentioned by Dr. Hill? I think these paradoxes arise, at least in part, through performing limiting operations in the wrong order. For example, the limit of $P.D.\ (x|y_0 < y < y_0 + \delta y_0)$ as δy_0 tends to zero is not necessarily equal to $P.D.\ (x|y = y_0)$ when $P\ (y = y_0) = 0$. (Here *P.D.* stands for "probability density"). To assume otherwise is equivalent to assuming that all Jacobians are equal to 1. Also, in the problem of the distribution on a sphere, there is a difference between a random great circle on a sphere rather than a great circle known to pass through a known fixed point (the North Pole). These two comments I believe remove the paradox from the example of the density on a sphere and the density on a longitude.

We all know that improper priors can sometimes be used if the limiting operations are performed in the right order. But one interesting example where an improper prior is definitely ruled out occurs in some work on Bayesian significance testing for multinomials and contingency tables (Good, 1965, 1967, 1976; Good and Crook, 1974; Crook and Good, 1979). In this work there is a Bayes factor $F(k)$ depending on a non-negative hyperparameter k such that the null hypothesis corresponds to $k = \infty$ and such that $F(k)$ tends to 1 when k tends to infinity. If a hyperprior $\Phi(k)$ is assumed for k such that $\int_0^\infty d\Phi(k)$ is divergent, then the resultant Bayes factor $F = \lim_{K \to \infty} \int_0^K F(k)\, d\Phi(k) / \int_0^K d\Phi(k) = 1$. In other words the evidence against the null hypothesis is completely annihilated by a prior that is "improper at infinity". Satisfactory results were obtained in the applications by using a proper prior that approximates the Jeffreys-Haldane improper prior of density $1/k$. The proper prior chosen for this purpose was a log-Cauchy with appropriate hyperhyperparameters.

Now consider Lester Dubins' problem (de Finetti, 1972, p. 205). An integer n has been selected by one of two procedures A or B. In procedure A the probability of a specific value of n is 2^{-n} $(n = 1, 2, \ldots)$, whereas in procedure B the probability is uniformly distributed. Thus

$$P\,(n \text{ is definable in less than } 10^{1000000} \text{ years } |\ B) = 0.$$

So if B is true we can never get evidence for it. (The universe is only about 10^{15} years old). But it was assumed that n *has* been defined. Therefore $P(B) = 0$. If we had originally judged that $P(B) = 0.5$, then we must change our minds in view of this additional thinking. I don't regard this as an inconsistency, in fact I have argued the value of what I call "dynamic probability". According to this theory we must admit that probabilities can change without new empirical information. See Good (1977).

Regarding the drunken-sailor problem, I don't see the advantage of explaining it in two dimensions rather than in one dimension. I think the problem then reduces to one discussed at the Waterloo conference on statistical inference, following a paper by Fraser.

L. PICCINATO *(University of Rome)*:

In principle I have some difficulty to understand fully what "complete ignorance" is, and I would prefer a slightly different approach. The model of professors Giron and Rios generalizes the usual model for decision problems in that it considers that we have not just one probability distribution on the states but that such law is known to belong to a given set. This generalization could be seen in a different way: the standard bayesian model is an ideal paradigm and it is surely useful to have some flexibility when we turn to practical applications (the case mentioned of several decision makers is an example). Therefore the perspective I would like to suggest is that of sensitivity analysis, or robustness, with respect to the choice of the prior.

The paper gives useful indications about how to proceed in this type of analysis. Anyway I am inclined to think that all Bayesians act sometimes as quasi-Bayesians in the sense of this paper: when we use conjugate distributions we are actually dealing with a problem which is in an intermediate position between total ignorance and a fully Bayesian approach where only one probability distribution on the states is requested. But in that case the use of classes of priors is only a matter of formal generality, which is attainable without any practical complication.

I think that these concepts about quasi-Bayesian procedures and the related mathematical aspects could be remarkably interesting in the framework of practical statistical analysis, I mean when K^* is suitably chosen in order to provide a better understanding of some decision problem. Some good examples of this kind were given e.g. by M. Skibinski and L. Cote (1963). The tools proposed by Giron and Rios could then usefully experimented along similar lines.

I suppose that professors Giron and Rios are substantially in agreement with me about the fact that the Bayesian approach provides an "ideal paradigm", in fact they essentially apply the Bayesian scheme in the rigourous classical way in correspondence with each element of K^*. This makes a remarkable difference with the approach by Skibinski and Cote who, unfortunately in my opinion, do not avoid integrations over the sample space. Let me recall about this that quasi-Bayesian procedures are sometimes imbedded in non-Bayesian frameworks, so that they can be misleading. For example, when two decision functions are compared in the standard non-Bayesian way (i.e. through the risk functions) one can find that decision d_1 is better than decision d_2 when θ belongs to a given subset Ω' of the state space Ω. It is then implicit that if you have a partial information that θ belongs to that subset (K^* could be a class of distributions with a support contained in Ω') you can say that d_1 is simply better than d_2, and hold that further information about θ is irrelevant. However this is not true, in general, also from an "objective" viewpoint (i.e. without using a specific prior) and in standard cases you could find experimental outcomes such that the terminal decision provided by d_1 is worse (in terms of losses) than the terminal decision provided by d_2, for every $\theta \epsilon \Omega'$ This depends of course on the fact that the risk functions are not admissible tools for a Bayesian analysis in post-experimental situations.

In conclusion let me say that I agree firmly with professors Giron and Rios in their attitude to relax usual assumptions without losing the basic aspects of the Bayesian approach, that is logical coherence.

A problem I would raise in connection with professor Hill's paper is the following: how to deal with statistical models if we must get rid with conglomerability? Of course the ground for accepting or refusing conglomerability (or complete additivity) is a logical one, and must be independent from the mentioned question. However, even if I agree that a logically sound approach needs essentially finite additivity only (so that complete additivity becomes a mathematical simplification to be used with care), it seems to me not irrelevant to seek what kind of implications this attitude has with respect to the standard statistical practice.

For "statistical model" I mean as usual a set of probability distributions on the space of possible outcomes, possibly indexed by a parameter whose actual value is unknown. Suppose that all conditional distributions are equal: then the value of the parameter seems irrelevant, at least from an intuitive viewpoint. In fact, if conglomerability holds, we can easily predict the future observations without knowing anything about that value. But, if conglomerability does not hold, it seems that something does not work well with our model, and our usual way of thinking, for example about the role and use of identifiability.

It is clear that a possible answer is that also statistical models must be handled with care, just as the assumption of complete additivity. For example, I think that this view is maintained by de Finetti, who dislikes such things as "statistical hypotheses" and so on; in fact he often suggests to deal only (or at least preferably) with well defined events, which could be actually falsified or verified; a specific quotation could be de Finetti (1971).

Nevertheless, as a statistician, I find that statistical models are quite useful, at least as a communication tool among the various kinds of researchers involved in a given joint work, and that often some relevant theoretical information can be imbedded in the model.

So, let us come back to the initial question: can we live with non-conglomerability (as professor Hill is proposing) without giving up with statistical models?

C. VILLEGAS *(Simon Fraser University):*

I will comment only on B.M. Hill's paper.

Finitely additive probability measures are conceptually important and are certainly useful in the foundations of probability and statistics. As a matter of fact I have used finitely additive probability measures in two papers (Villegas 1964, 1967). But for technical reasons it is usually better to assume countable additivity.

In recent years increased interest has been shown in the use of betting schemes for the analysis of statistical inferences. In those betting schemes the statistician plays the role of a bookie that posts odds on a family of events, and has to withstand the bets of a gambler. The odds posted by the statistician are said to be coherent (relatively to the betting scheme) if the statistician cannot be made a sure loser. In this betting context support for countable additivity comes from Theorem 6 of Heath and Sudderth (1972). Roughly speaking, the theorem says that if, in the absence of data, the gambler is allowed to make countably many bets, then the posted odds will be coherent if and only if they are based on a countably additive probability measure.

Similar results hold when data are available. Thus, Corollary 1 of Heath and Sudderth (1978) says that, if the gambler is allowed to make only a finite number of bets, then the posted conditional odds will be coherent if and only if they are based on a posterior distribution corresponding to a finitely additive, proper prior.

These results can be extended to obtain a justification for countably additive proper priors. Thus, in a future paper, I will prove that, if the gambler is allowed to make countably many bets, then the posted conditional odds are coherent if and only if they are based on a posterior distribution which corresponds to a proper, countably additive prior.

A new conditional frequency interpretation of statistical inferences has been offered in Villegas (1977a). Future repetitions used in frequency interpretations are not real but hypothetical or simulated, and they should be considered only as a means for learning from the data. In Villegas (1977a) it is argued that better inferences may be obtained if only future hypothetical samples similar to the actual data are considered, because in this way the noise may hopefully be reduced, and we may get a better picture of what the actual sample has to say about the population.

Looking only at future samples which are similar to the actual one means conditioning on the future hypothetical sampling belonging to a compact set. Within the context of a betting scheme this means that all bets are off if the observation does not belong to a compact set.

Using this form of conditioning, the results of Heath and Sudderth can be extended to obtain justifications for improper priors. Thus, in a future paper, I will prove that, if the gambler is allowed to make countably many bets, but all bets are off when the observation is outside a compact set (which may be chosen by the gambler), then the posted conditional odds are coherent if and only if they are based on a posterior distribution which corresponds to a possibly improper, countably additive prior.

Objections against the use of improper priors have been raised also from the point of view of admissibility. Thus, in the estimation of a location parameter, the Bayes estimators based on a uniform prior may be inadmissible (Stein, 1956), However, results will in general be different if we condition on the observed value belonging to a compact set. Then the risks become conditional risks, and a new concept of conditional admissibility emerges. In a future paper it will be shown that it is not difficult to modify C.R. Blyth's (1951) proof of admissibility of Bayes estimators based on improper priors are conditionally admissible in the above mentioned sense. And, according to the new frequency interpretation of Villegas (1977a), this is all that is needed in statistical inference.

It should be recognized that there are two lines of development for Bayesian statistics: one is the personalistic line, based on personal, subjective priors, and the other is the logical probability line, based on logical priors that represent ignorance. The second line is not so well developed as the first one, but some progress has already been made (Villegas, 1977b).

Logical priors are usually invariant under a given group. Therefore they are not only relative to a given model, but even more, they are relative to a group that is given as an integral part of the model. Fraser's structural models are useful from a logical probability viewpoint. Stone's example becomes a structural model if the group with two generators is considered as an integral part of the model. In that case the logical prior is the uniform prior. But the story of the lady and the sailor brings other considerations which favor the selection of the other prior. Since the likelihood principle ignores the possibility that a group may be given as an integral part of a model, it is not valid in a logical probability approach to statistical inference.

J.M. DICKEY *(University College of Wales Aberystwyth):*

I find the paper by Professors Giron and Rios intriguing, especially the idea of working with "extremal" posterior distributions to surround, so to speak, the coherent inferences of persons whose prior distributions lie within a range of distributions. This harmonizes closely with my idea of "scientific reporting" as a reporting of the prior-to-posterior transformation over a class of prior distributions conceived as containing the reasonable uncertainties of a population of scientists (Dickey, 1973). Various graphical methods are available for reporting such a distribution-valued functional. Bounding methods are also proposed in both papers.

The idea of Giron and Rios seems simple and straightforward, and in view of the long story of statistical theorists saying they could not know their prior distributions, one would have expected this idea to have developed much earlier. The authors have done a great service in carefully setting out the theory. I look forward to seeing more applications.

An obvious direction of generalization which may interest the authors is to replace the set K^* of permitted, equally acceptable, prior distributions by a new distribution of distributions, an expression of uncertainty concerning uncertainty. This could be used to generate sets K^*, for example, by setting thresholds on some density for the new distribution in function space. My own paper in this meeting investigates the form such a distribution might take and its use in the problem of assessing (choosing) a subjective probability distribution. See also Dickey and Freeman (1975). There are, of course, logical difficulties with the meaning of such a second-order belief distribution, and in both our settings one would need to resist the temptation to marginalize by taking the second-order average of first-order beliefs.

Finally, I should like to complain that the term "agreement set" for K^* or its convex closure could be misleading. Presumably, the decision makers agree in having their opinions fall in the set. But then they **disagree** on which distribution is appropriate *within the set.*

There are many diverse issues raised in Proffesor Hill's paper. The main point for me is that he argues with De Finetti in favour of merely finite additivity, and consequent nonconglomerability. In the sphere example this would mean that *all* the great circles through the poles could have uniform distributions within a circle, while the two-dimensional probability on the sphere could also be uniform. This conflicts with the conditional distribution that would be obtained by a limiting argument conditioning on an observed small interval of longitudes.

I am grateful to Professor Hill for personal conversations in which he informed me that his issue in the sphere example is not the same issue as brought forward by Kolmogorov (1933, Ch. V, Sec. 2). Kolmogorov cites Borel for what I have called the Borel-Kolmogorov nonuniqueness, whereby a conditional distribution obtained in the usual way from a joint density will depend on the conditioning variable used to define the conditioning event, rather than just on the conditioning event itself. In the sphere example, a different experiment which slices the earth by parallel planes will produce uniform distributions within the circles produced.

Apparently, Hill is not thinking of any experiment at all when he asks for the distribution within a great circle, but wants to base a conditional distribution on the purely logical statement that a particular great circle obtains. He wants finite additivity "in part for purely logical reasons". He also claims to need it for practical reasons, since "one will often find it advisable to make approximations using infinite models".

I simply do not understand the practical need for merely finite additivity. When I make approximations to finitistic situations using infinite models I shall not restrict myself to using only a few logical statements to obtain a mathematical model. I shall look at the real-world problem and the real uncertainties involved. For example, just because some exercises in textbooks fail to give information distinguishing between equal-length intervals would not be enough to tempt me in a real-world problem to use a uniform pseudodensity over the whole real line. It seems to me that countable additivity, conglomerability, and proper integrable distributions enable us to treat real problems realistically, without worrying that the mathematics itself will deal us an unpleasant surprise. I should like to hear further about the practical issues. Mervyn Stone's lazy-Bayesian examples over the years have only served to warn us against nonintegrable distributions, which were already ruled out by the axioms of coherent behaviour.

M.H. DEGROOT *(Carnegie-Mellon University):*

In the paper by Rios and Giron, partial information about a prior distribution is represented by simply dividing all distributions in Ω^* into a set K^* of possible prior distributions and the complementary set of impossible prior distributions. Wouldn't it be more reasonable to assign probabilities to the distributions in Ω^* ; i.e., to assign a probability distribution P^{**} to the set Ω^* . In turn, one might then assign a distribution P^{***} to the set Ω^{**} of all distributions P^{**} , etc. In brief, why not develop a hierarchical model?

D.A.S. FRASER *(University of Toronto):*

I wish to discuss three points connected with Professor Hill's paper: how the Stone example provides a strong counter example to the Strong Likelihood Principle; how the modelling of the internal variable of the Stone example leads to the overriding probability statements; and how information concerning a realization from such an internal variable must satisfy certain requirements as to how it was produced in order to be acceptable for probability calculations.

The Stone example A has seemed to me to be a very striking counter example to the Strong Likelihood Principle. Professor Hill has doubts and discusses the distinctions between the full parameter and two interesting component parameters. The full parameter for the model is $\theta = p$, the path from the origin to the treasure; a derived parameter of interest is $\theta_1 = \theta_1(\theta) = \vec{x}$, the last directed segment of p; a further derived parameter of interest is $\theta_2 = \theta_2(\theta) = x$, the end point of the path p. These parameters are not the same and yet, given a data-point $\hat{p}$ (the path to the sailor), the possible values for them fall into a one-one equivalence. The observed likelihood function is a function of the full parameter θ; as presented it is not a likelihood for either component parameter but does of course provide information concerning each. The full parameter space is $\Omega = \{p\}$, the free group on two generators.

A salient feature of the Stone example is the striking contrast between the following two results: the likelihood function from data assigns equal likelihood ($\frac{1}{4}$) to each of four possible paths to the treasure; direct probability arguments based on an internal variable put an operational 3/4 probability on a preferred one of the four possible paths. Thus, likelihood says the four possibilities are on a par one-with-another, whereas an internal variable nominates one of the four possibilities as a 75% favourite. The example seems to make clear that likelihood does not contain all the needed information.

Perhaps some further details can add emphasis to this result. For the Strong Likelihood Principle my own preference is a prescription in the following form: from a statistical investigation use only the observed likelihood function. An alternative form closer to that proposed by Birnbaum is the following: if the likelihood function from a first model + data-point is the same as the likelihood function from a second model + data-point then the inferences should be the same in the two cases. For this we note that the likelihood function is a nonnegative function on the parameter space Ω left indeterminate to a positive multiplicative constant; that is, it is a positive ray from the origin in the vector space R^{Ω}. The equality, then, of two likelihood functions requires the same parameter space Ω and the same ray in R^{Ω}.

Is the probability imbalance and the constant likelihood on four parameter points, a necessary consequence of the unusual parameter space? Or could we find another model + data-point that yields an identical likelihood function but with a different probability imbalance or more simply with say symmetry on the four possible parameter values? We examine this latter possibility.

For this suppose we start with some particular likelihood function obtained from the Stone example with a data-point; let $\hat{p}_0$ be the data point and $\theta^1, \theta^2, \theta^3, \theta^4$ be the four possible parameter values consistent with $\hat{p}_0$. For a second model we take the same parameter space Ω, the same sample space $S = \Omega$, and the following very special probability structure:

$$
\begin{array}{ll}
P(\hat{p}_0|\theta^i) = \frac{1}{4}, p(e|\theta^i) = \frac{3}{4} & i = 1,\ldots,4 \\
P(e|\hat{p}) = 1 & \\
P(\theta|\theta) = \frac{1}{4}, P(e|\theta) = \frac{3}{4} & \theta \neq \theta^i, \theta \neq \hat{p}
\end{array}
$$

where e is the identity element. The likelihood function from the sample point $\hat{p}_0$ is the same as that from the Stone example and yet the model treats the four parameter values symmetrically. This provides the formal contradiction to the Strong Likelihood Principle.

Clearly the likelihood function alone is not enough. Of course many statisticians do not accept the Strong Likelihood Principle, usually on the good grounds that many fruitful statistical results are available outside the Principle. The Stone example however is direct: the likelihood function alone omits an essential probability property.

The Stone example contains a primary random system - the spinning of the woman at the end of the taut thread. Based on this process, there is an overriding 3/4 probability that the path is extended, and correspondingly an overriding 3/4 probability that the last path segment comes from the treasure. This seems to provide the motivation for Stone's "classical statistician" although details are not given. A formal version of the preceding appears in my Comments on the Stone paper but was sidestepped in Stone's elusive rejoinder. The recognition of the fundamental importance of primary or internal random systems seems long overdue in contrast with the intensive activity in some areas of contemporary statistics.

Prof. Hill also considers the system in which a point is selected uniformly on the surface of a sphere with a designated north and south pole; an investigator is given the exact longitude of the point. Prof. Hill seems to show preference for a uniform distribution for the point on the given great circle of longitude. This is in conflict with a basic probability position, both classical and Bayesian, that marginal and conditional probabilities go together to give joint probabilities. For we note that the standard conditional distribution given that longitude equals the recorded value has density proportional to the cosine of the latitude.

What is the key element in the preceding conflict? We have a situation where there is information concerning a realization from a random system, and yet the information does not fully identify the realization. Discussions of conditional probability show that we need to know not only the information as to possible values for a realization **but also** how that information was produced; see for example Fraser (1976, Ch. 4), Fraser and Brenner (1979).

Most discussions of conditional probability overlook the need to know how the information is produced concerning the possible values for the concealed realization. Without it, contradictions are obtained and various "paradoxes" are to be found in the literature. Information without knowledge concerning its production does not support probabilities. This is a very fundamental argument against the Bayesian position.

S. FRENCH *(University of Manchester):*

I wish to comment upon Girón and Ríos's paper. First, a few points of a technical nature. The authors have to use topological properties of Ω and ideas of continuity in case (a) of their theory. I wonder if these assumptions can be weakened by using the approach of Krantz et al. (1971). These latter authors have avoided the use of topological assumptions in their measurement systems instead relying on weaker solvability conditions applied to the underlying qualitative orders. Perhaps Girón and Ríos could generalise their results similarly.

Early in their Paper, Girón and Ríos discuss partial orders derived from convex cones in $\mathbb{R}^n$. I wonder if they have seen the recent work of Hartley (1978). His approach seems to give the weakest set of conditions available for playing with such orders. Also for a practical illustration of the use of such cone-orders in the sensitivity analysis of a decision problem, the authors have referred to Fishburn (1964). His paper (1965) in Operations Research is also of relevance and, perhaps, easier to find.

Turning now to what I believe to be a more important question. The authors consider a decision maker who knows his utility function perfectly and his subjective probabilities imperfectly. Is this a reasonable model? It says essentially that he can locate for each possible consequence an exactly equivalent gamble based upon some auxillary experiment. Is it feasible to suggest that he can do this, yet be unable to locate a gamble based on the auxillary experiment equivalent to a gamble based upon an unknown state of nature? The problem of measuring subjective probability is just as easy, or difficult, as that of measuring utility. In terms of axiom systems my point is this. In assuming the existence of a utility function $u(\cdot)$ the authors are hiding under their decision space another decision space in which the ordering of decision rules is complete.

Finally, since I see the primary use of this theory to be in the area of sensitivity analysis, perhaps the following suggestion is appropriate. I have seen papers in which, as here, the utilities are known and the probabilities only partially known and also papers in which the probabilities are known and the utilities partially known. I wonder if duality theorems of mathematical programming can give us a means of allowing both quantities to be partially known? Perhaps the authors know of a reference in this area.

D.V. LINDLEY *(University College London):*

I have a brief comment on the paper by Girón and Ríos. How does a partially ignorant person act? Bayesian decision theory is a recipe for the selection of a single act: Bayesian inference provides all the information about the unknowns in the problem needed to select the act. The authors' theory ends with a class of acts: if this class contains more than one member, how is a unique act to be selected in cases where no more data is available? A possible application of this theory is to multiple decision problems where several opinions are present, but again there is the difficulty of the choice of a single act.

Turning now to Hill's paper, Kolmogoroff (1933 Ch. 5), makes the point that conditional probability is either defined with respect to an event of non-zero probability, or for a random variable $x(w)$ defined over a space of values of w, and not for the single event $x(w) = x_0$ when this has probability zero. My understanding is that Kolmogoroff would want to know what random variable gave longtitude 30; was it longtitude, or was it some other variable? This seems right to me and I'd welcome Hill's comments on this. It constrasts with the likelihood principle since it requires knowing not just that the longtitude was 30 but what other values (like 25) one might have had. What are the "gaping holes" - mentioned in the first paragraph - in a sigma-additive theory using proper distributions?

REPLY TO THE DISCUSSION

F.J. GIRON *(Universidad de Malaga)* and S. RIOS *(Universidad de Madrid):*

We would like to start by paraphrasing Dempster, quoted by Bernardo (1979): "In the area of statistical inference, there must be little that any one has thought about that Dr. Good has not written about, to the point that a computerized information retrieval system would be very helpful to scholars in the area".

Our paper does not intend to be a historical paper nor a paper on the history of partially ordered probabilities, and explicit reference to previous ideas on the subject are mentioned in section 1.

With respect to the priority claimed by Professor Good, it is worthwhile mentioning here that the idea of approximating sharp probabilities by means of an interval is to be found in an early paper by Fréchet. Unfortunately we have not been able to trace back the appropriate reference thought it might be found in Econometrica. To what extent early ideas influence a theory is always a controversial subject. As an example some french authors and others refer to the Kolmogorov axioms as the Fréchet-Kolmogorov axiomatic set up.

We agree with Professor Piccinato that the Bayesian approach is the "ideal paradigm". Yet to contemplate the quasi-Bayesian theory merely as a sensitivity analysis approach is, we believe, to focus just on a particular aspect of the model. Its interest resides in that the hypothesis of the model are more general than that of the Bayesian model; more mathematically tractable than other former approaches (the one mentioned by Professor Piccinato of Skibinski and Cote (1963) could be an example); and above all in the main theorem that establishes an equivalence between the ideas of partial ordering of decision rules and partial information in terms of probability measures. On the other hand, the interpretation of the theory from the point of view of sensitivity analysis also stressed by Dr. French in his contribution to the discussion, allows for a unified and systematic treatment of the problem of sensitivity analysis in Bayesian decision making.

With respect to the problem of non-admissibility of quasi-bayesian procedures that Professor Piccinato mentions nearly at the end of his contribution, the situation here is exactly the same as in the Bayesian case. Problems of admissibility in post-experimental situations depend on three facts: 1st, prior partial information may be incompatible with some experimental outcomes; 2nd, the support of distributions of K^* may be a proper subset Ω' of Ω, thus discarding some states of Nature; 3rd, the judicious use of Fubini's theorem.

We are grateful to Professor Dickey for his comments and, like him, we would also like to see more applications of the theory. We have taken up his complaint and have change the term "agreement set" into the more innocuous term, and we believe it more apt, "feasible set".

The generalization suggested by Professor Dickey, which is also pointed out by De Groot in his contribution, of developing a hierarchical model seems interesting, specially the idea of setting thresholds on some distribution of distributions (the second stage in the hierarchical model) to generate sets K^* of first-orders beliefs. This idea is also closely related to the paper by De Robertis and Hartigan (submitted for publication to the Annals of Statistics) about ranges of measures as an expression of partial ignorance.

Professor De Groot's suggestion of developing a hierarchical model is discussed at length in the paper by Good at this conference. However as he presents the hierarchical model we would have in the first stage a complete ordering given by the probability measure P^{**}. In the second stage, we would now have as new states of Nature the set of all probability measures on Ω^*, that is Ω^{**}, on which a new distribution P^{***}, could be assigned, and so on; so that this would drive to a complete ordering of decision rules by marginalizing on succesive stages unless in any of the stages the probabilities assigned were partially ordered (cf. Good, p. 7, line 12 of his revised manuscript) and thus the final ordering of decision rules would only be partial.

Our paper is an attempt to characterize these partial orderings which, of course, can be embeded in a hierarchical model, one of the stages of which at least corresponds to partially ordered probabilities.

Dr. French suggests a generalization of our paper by using the approach of Krantz et al. (1971). We believe this program can be carried out along their lines. Another possible generalization of the results of our paper for partial comparative probabilities, that also takes into account the role of experimentation, could be based on the works of Fine (1971, 1973). Yet we want to point out two facts: 1st, in the Krantz et al. approach the subjective probability derived is finitely additive as in case (b) of our paper, in which the only requirement is the existence of a bounded utility function; 2nd, the topological assumptions of case (a) guarantee the σ-additivity of probability measures of set K^* and neither compacness of Ω nor continuity of acts can be dropped if one is seeking for σ-additive subjective probability measures. Further, this allows for a parallel and systematic treatment of cases (a) and (b) and renders the proofs of main theorems almost trivial by using the topological dual of spaces $C(\Omega)$ and $B(\Omega)$, respectively.

Unfortunaltely the paper by Hartley (1978) French mentions has not reached our hands at the time of writing the rejoinders.

We are in agreement with Dr. French when he says that our model is not reasonable because it takes for granted that utilities are perfectly known and, in practice, both quantities, probabilities and utilities, are only partially know. However we know of no duality theorem of mathematical programming that can accomodate the case when both quantities are partially known, although we think this to be a very important issue in practical decision making.

The question Professor Lindley raises is a key one; namely, how does a partially decision-maker act? The answer is in the premises of the theory, precisely in the formulation of Axiom 1. If a partially ignorant person has only a limited amount of information, then he selects a class of non-dominated acts such that it is worth while betting on these acts against other acts. Usually, this class contains more than one act, and then it is not clear how a single act is to be selected. A possibility would be to randomize among these acts, but this would be equivalent to consider a hierarchical model and this, in turn, is equivalent to having your decisions linearly ordered.

On the other hand, Bayesian decision theory may also lead to a class of acts (when several decisions attain the same Bayes risk) and then it is not also clear how to randomize.

In short, if one is partially ignorant one cannot expect to be able to linearly order the set of possible decisions.

Quasi-Bayesian theory takes into account the possibility of partial-instead of total-information thus generalizing Bayesian theory. Then, it is proven that such a hypothesis is intimately related to partial ordering of decisions as opposed to the complete ordering of decisions in Bayesian theory. Which is more plausible is a question of applicability and even of taste.

B.M. HILL *(University of Utah and University of Michigan):*

I would like to thank all of the discussants for their comments. Before responding to individual discussants it may be helpful to make some general remarks. The primary purpose of my article was to focus attention on the axioms for Bayesian inference and decision theory. The de Finetti axioms are weaker than others in that they allow finitely additive distributions and non-conglomerability. It is hard to imagine satisfactory axioms for quantitative probability that are still weaker than those of de Finetti, and failure to abide by axioms 1 and 2 can subject one to sure loss. Should, however, these axioms be strengthened? Should, for example, one require that decision procedures be extended admissible, or perhaps even admissible. If there are serious arguments so to strengthen the de Finetti axioms, then there should exist telling examples clearly demonstrating the shortcomings of the finitely additive approach. The examples that I chose were those that seemed most clearly to suggest possible shortcomings , and I attempted to determine just how serious a case could be made to strengthen the axioms. Thus in Mervyn Stone's example, I think most of us will prefer the Bayesian solution based upon a uniform prior distribution for N, whether this is taken literally or as an approximation using proper prior distributions. The de Finetti axioms, however, do not exclude the finitely additive prior distribution $\tilde{q}(\cdot)$ that leads to the Stoned Bayesian Posterior. So it seems natural to ask exactly what ill consequences will occur if one were to use this prior distribution. Stone suggested that over a long sequence of repetitions of the experiment the Stoned Bayesian would get the treasure less frequently than someone who used the confidence solution. My discussion of the sphere example was meant to suggest why his argument is not very convincing even within the frequentist theory. For it is circular. Only if you have already rejected finite additivity and non-conglomerability does the argument suggest an unambiguous frequency for obtaining the treasure.

Now let me turn to the individual discussants. Professor Good suggests that the paradoxes (if such they be) arise from incorrect limiting arguments. I do not think so. Indeed, there are no limiting arguments in my article, and I tried to avert such a misinterpretation by conditioning upon an *exact* great circle. Admittedly this is an idealization for real world problems. But so conditioned the problem is still logically meaningful, analogous idealizations are commonly made in statistics, and there can easily arise situations where the appropriate conditioning event is not specified, i.e., we are not told whether the measurement process restricts us to the region between two parallel planes, or between intersecting planes through the poles, or still other regions. (Such sensitivity to the precise form of the conditioning event is still another reason to argue for the freedom of the finitely additive approach). Would Professor Good, along with Professor Fraser, simply refuse to discuss the question in the absence of such information? Professor Good then refers to the distinction between a random great circle on a sphere and a great circle known to pass through a fixed point (See also my footnote # 3). He should then be able to point to the ill consequences from taking the point as uniform on the great circle in the latter case. But I suspect that he will only be able to demonstrate such consequences if he has already assumed countable additivity and consequently also conglomerability. With regard to Professor Good's discussion of the Dubin's problem, I find his argument that $P(B)=0$ even less convincing than my own that $Pr\{T_2\}=0$ in Example 3. First of all the age of the universe is not so terribly well known as he implies. Would Professor Good be greatly surprised if by the year 2,079 some new theory suggested that the age should be revised upwards to 10^{25} years, or whatever? Secondly, I am concerned with his emphasis on "definability". Suppose we are discussing the number of elementary subatomic particles in the universe, and for the sake of argument assume that there is a well-defined number. Then although under hypothesis B it will probably take awfully long to "define" this number, the number has been assumed to exist, and the finitely additive uniform distribution (at least in the upper tail) may represent ones' opinions much more adequately than any countably additive distribution. What if, for example, one simply cannot name a number such that the probability to the right of that number is less than 10^{-100}?

I find Professor Good's discussion of "dynamic probability" intriguing. But I doubt that it is relevant to the Dubin's problem or Example 3. The reason for my doubt is that the alteration in $Pr\{B\}$ or $Pr\{T_2\}$ that he suggests would be made merely to avoid non-conglomerability, without having advanced any serious argument as to the need for conglomerability. Finally, I was sorry that Professor Good did not choose to

discuss the drunken-sailor problem. Although the one dimensional version has much in common with it, there are certainly real differences between the two versions, for example, the non-amenable free group on two generators, and in particular the finitely additive analysis of the problem in two dimensions would appear to be new.

Professor Dickey questions the practical need for merely finitely additive distributions. I think all three of the examples I discussed suggest such a need. In the drunken-sailor example Professor Dickey presumably would object to the uniform finitely additive distribution on N, and at best would view it only as an approximation for a proper countably additive distribution. Even so, is it not sometimes useful to have available such a simple approximation, rather than to labor over the fine details of ones prior distributions in a situation where there is little to be gained from such labor? Similarly for the problem on the sphere. What if Professor Dickey does not have available all the real-world information he would like, so that the shape of the region delimited by the actual measurement process is not known. Keeping in mind the possibility of parallel hyperplanes, would he exclude the uniform distribution on a great circle, even as an approximation? Would he simply ignore the problem, as so many non-Bayesians do with regard to any problem that doesn't fit into a neat Kolmogorov-frequentistic mold? Finally, improper prior distributions can often be given a finitely additive interpretation, so that they are in fact consistent with the de Finetti axioms for coherent behaviour. (See my footnote n° 4.)

Professors Dickey and Lindley both point out that in the Kolmogorov approach it is not sufficient to know the conditioning event, and that one also needs information regarding the conditioning variable used to obtain that event, at least when the event has probability zero. This is true, and seems to me to cast doubt upon the approach itself. As I argued above, does this mean one should say nothing when such information about the variable is not available? My notion of uniformity on the surface of the sphere includes not only the evaluation of probabilities as proportional to surface area for sets that have surface area, but also the notion that conditional upon the point being in any specified finite sets of points, all such points are equally likely, and conditional upon a great circle, probability is proportional to arc length. This strong notion of uniformity is not possible in the Kolmogorov approach, but is compatible with the de Finetti axioms. Why should such an opinion be excluded? The contrast between the Kolmogorov approach and the likelihood principle is itself one of the gaping holes. Conventional statistical models often assume the data to have probability zero, and within the model Bayesians are forced to consider their probabilities conditional upon an isolated event of zero probability, although Kolmogorov (1933, p. 51) wishes to exclude precisely this situation. Of course one can take refuge in a finitistic approach, but then we lose the advantages in simplicity that we obtain with conventional models. I think the situation is somewhat akin to that with regard to stopping rules and the likelihood principle. A conventional non-Bayesian analysis is not really possible without knowledge of the stopping rule, and since we rarely if ever know the true stopping rule, a conventional analysis could at best yield only certain inequalities. In the same way, a conventional Bayesian analysis in the Kolmogorov system is only possible if one knows the conditioning variables, and I submit that in most applications they too are unknown. But we will nonetheless draw inference and make decisions. I believe that the arguments against such an approach are circular. They have force only if one has already accepted countable additivity.

Professor Villegas has an interesting alternative way to deal with inadmissibility, but it does not seem appropriate to discuss this here.

Now let me turn to Professor Fraser's comments. Despite my very best efforts Professor Fraser still regards the Stone example as convincing evidence against the likelihood principle. I cannot agree. First of all, the only kind of likelihood principle that can have any credibility at all is one compatible with Bayesian inference. For even a non-Bayesian would have to reject a version of the likelihood principle that was not compatible with Bayesian inference whenever he thought that the prior distribution had a frequency interpretation. This in turn implies that a data-dependent transformation of the original parameter must be excluded as evidence against the likelihood principle, since the transformed parameter would have a different "prior" distribution than the original parameter, as I hope my discussion of E and $\dot{E}$ makes clear. Professor Fraser apparently now accepts this but offers still another experiment to provide a "formal contradiction to the Strong Likelihood Principle" (nearly the same as my likelihood principle). In order for his new experiment to make sense we must assume that the new experiment consists in first performing the original experiment to obtain his data $\hat{p}_0$ (my $\hat{p}$), and then performing some additional experiment to generate his new likelihood function. (Note that this must be done for all possible $\hat{p}_0$, not just a particular realization). Even if he is correct that the modified experiment yields the same likelihood function as the original experiment the argument loses its force because whatever asymmetry is involved in the original experiment must then be reflected in Fraser's modification. But his purpose was to treat the four parameter values symmetrically.

Professor Fraser also argues against the likelihood principle on the grounds that it counters many "fruitful statistical results" It is of course counter to conventional significance testing, but Bayesians are hardly alone in regarding such tests with a great deal of skepticism.

Finally, Professor Fraser discusses the need to know how information is produced, as was raised by Professor Lindley and discussed above in my reply. This is presumably a much more fundamental issue for Professor Fraser than for Professor Lindley, and is at the root of much criticism of the Bayesian approach, dating back at least to Venn. Thus Professor Fraser presumably would have us do nothing without such knowledge, and also without knowledge of stopping times, etc. This perhaps restricts the applications of statistics to the empty set. I would also ask Professor Fraser exactly how we are to discriminate between the various forms of knowledge, i.e., between knowledge that can be (in his sense) validly represented by a probability distribution, and opinions that cannot be so represented?

Professor Piccinato raises some intriguing questions regarding the use of conventional statistical models. As I see it finite additivity and non-conglomerability offer us some additional freedom in the probabilistic expression of our knowledge. In some applications it will be important to take advantage of that freedom, and in some it will not. As in my reply to Professors Dickey and Lindley, I think that in the sphere example it is important not to force oneself into the Kolmogorov mold, at least not without careful consideration as to the knowledge that one wishes to express. But I do not think there is anything incompatible between the careful use of conventional statistical models and the de Finetti theory. It is true that conventional parameters can often be dispensed with, as for example in an exchangeable sequence of zero-one variables, and where this is possible it seems preferable to do so rather than to invent artificial parameters. (In Hill, (1969), it is shown how conventional linear models can also be dealt with in this way). But on the other hand there are many situations which cannot as yet be handled satisfactorily in terms of the observable variables, and parametric models offer a convenient flexible way of dealing with such situations. In any case it is not a question of incompatibility, but merely of seeing things in another light. Finally the question as to the case where the conditional distributions are the same, and so as Professor Piccinato suggests, the parameter might seem to be irrelevant, is indeed a paradox of non-conglomerability. But despite the intuitive plausibility of merely dispensing with the parameter, perhaps we should recall that we must have had some reason to view the situation as non-conglomerable in the first place, and then to choose as best we can between the conflicting intuitions.

REFERENCES IN THE DISCUSSION

BERNARDO, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. B* **41**, 113-147.

BLYTH, C.R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* **22**, 22-42.

BRENNER, D. and FRASER, D.A.S. (1979). Conditional probability and the resolution of statistical models. *Statistische Hefte*. (to appear).

CROOK, J.F. and GOOD, I.J. (1979). Part II of Good (1976). *Ann. Statist.* **20**, 148-159.

DE FINNETI, B. (1971). Probabilitá de una teoria e probabilitá dei fatti. *Studi di Probabilitá Statistica e Ricerca Operativa*, 86-101. Universitá di Roma: Istituto di Calcolo delle Probabilitá.

— (1972). *Probability, Induction, and Statistics*. New York: Wiley.

DeROBERTIS, L. and HARTIGAN, J.A. (1979). Ranges of prior measures. *Tech. Rep.* Yale University

DICKEY, J.M. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *J. Roy. Statist. Soc. B* **35**, 285-305.

DICKEY, J.M. and FREEMAN, P.R. (1975). Population-distributed personal probabilities. *J. Amer. Statist. Assoc.* **70**, 362-364.

FINE, T.L. (1971). Rational decision making with comparative probability. *Proc. IEEE Conf. on Decision and Control*. 355-356.

FISHBURN, P.C. (1964). *Decision and Value Theory*. New York: Wiley.

— (1965). Analysis of decisions with incomplete knowledge of probabilities. *Operations Research* **13**, 217-237.

FRASER, D.A.S. (1976). *Probability and Statistics, Theory and Applications*. Toronto: University of Toronto. Textbook Store.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence.* New York: Hafners.

— (1952). Rational decisions. *J. Roy. Statist. Soc. B.* **14**, 107-114.

— (1962). Subjective probability as the measure of a non-measurable set. *Logic, Methodology, and Philosophy of Science: Proc. of the 1960 International Congress* (Stanford), 319-329.

— (1965). *The Estimation of Probabilities.* Cambridge, Mass.: M.I.T. Press.

— (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. B.* **29**, 399-431.

— (1976a). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**, 1159-1189.

— (1976b). The Bayesian influence, or how to sweep subjectivism under the carpet. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (Hooker, C.A. & Harper, W., eds.) Vol. 2, 125-174, Dordrecht: D. Reidel.

— (1977). Dynamic probability, computer chess, and the measurement of knowledge. In *Machine Intelligence* **8** (Elcock, E.W. and Michie, D., eds.) 139-150, New York: Wiley.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/non Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

HARTLEY, R. (1978). On cone-efficiency, cone-convexity and cone-compactness. *SIAM J. Appl. Math.* **37**, 211-222.

HEATH, D.C. and SUDDERTH, W.D. (1972). On a theorem of De Finetti, oddsmaking, and game theory. *Ann. Math. Statist.* **43**, 2072-2077.

— (1978). On finitely additive priors. *Ann. Statist.* **6**, 333-345.

HILL, B.M. (1969). Foundations for the theory of least squares. *J. Roy. Statist. Soc. B* **31**, 89-97.

KEYNES, J.M. (1921). *A Treatise on Probability.* London: MacMillan.

KOLMOGOROFF, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer. English translation as *Foundations of the Theory of Probability.* New York: Chelsea, (1950).

KOOPMAN, B.O. (1940a). The basis of probability. *Bull. Amer. Math. Soc.* **46**, 763-764.

— (1940b). The axioms and algebra of intuitive probability. *Ann. Math.* **41**, 269-292.

KRANTZ, D.H. & *et. al.* (1971). *Foundations of Measurement, Vol. I.* New York: Academic Press.

SKIBINSKI, M. and COTE, L. (1963). On the inadmissibility of some standard estimates in the presence of prior information. *Ann. Math. Statist.* **34**, 539-548.

SMITH, C.A.B. (1961). Consistency in statistical inference and decision. *J. Roy. Statist. Soc. B* **23**, 1-25.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp.* **1**. 197-206.

VILLEGAS, C. (1964). On qualitative probability σ-algebras. *Ann. Math. Statist.* **35**, 1787-1796.

— (1967). On qualitative probability. *Amer. Math. Month.* **74**, 661-669.

— (1977a). Inner statistical inference. *J. Amer. Statist. Assoc.* **72**, 453-458.

— (1977b). On the representation of ignorance. *J. Amer. Statist. Assoc.* **72**, 651-654.

Trab. Estadist. (1980)
Vol. **31** pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 67-142 Valencia, Spain: University Press

2. Sequential learning, discontinuities and changes

INVITED PAPERS

MAKOV, U.E. *(Chelsea College, London)*
Approximations of unsupervised Bayes learning procedures.

SMITH, A.F.M. *(University of Nottingham)*
Change-Point problems: approaches and applications.

HARRISON, P.J. and SMITH, J.Q. *(University of Warwick)*
Discontinuity, decision and conflict.

DISCUSSANTS

FIENBERG, S.E. *(University of Minnesota)*
BERNARDO, J.M. *(Universidad de Valencia)*
BROWN, P.J. *(Imperial College, London)*
DAWID, A.P. *(The City University, London)*
DICKEY, J.M. *(University College of Wales)*
KADANE, J.B. *(Carnegie-Mellon University)*
LEONARD, T. *(University of Warwick)*

REPLY TO THE DISCUSSION

Approximations of unsupervised Bayes learning procedures

U.E. MAKOV
Chelsea College, London

SUMMARY

Computational constrains often limit the practical applicability of coherent Bayes solutions to unsupervised sequential learning problems. These problems arise when attempts are made to learn about parameters on the basic of unclassified observations, each stemming from any one of k classes ($k \geq 2$).

In this paper, the difficulties of the Bayes procedure will be discussed and existing approximate learning procedures will be reviewed for broad types of problems involving mixtures of probability densities. In particular a quasi-Bayes approximate learning procedure will be motivated and defined and its convergence properties will be reported for several special cases.

Keywords: SEQUENTIAL; CLASSIFICATION; ESTIMATION; BAYESIAN; QUASI-BAYES; STOCHASTIC; APPROXIMATION; IDENTIFIABLE MIXTURES.

1. INTRODUCTION

Problems of unsupervised learning arise when attempts are made to learn about parameters on the basis of sequential unclassified observations each stemming from any of k classes ($k \geq 2$). General discussions of such problems in the contexts of Pattern Recognition and Signal Detection are given in Fu (1968), Patrick (1972), Young and Calvert (1974) and references there cited.

In this paper, we shall consider the following special cases. (See a survey in Ho and Agrawala (1968), for a discussion of these and other cases).

Case A. The probabilities, $\pi_1, \ldots, \pi_k$ that an observation belongs to class H_i, $i = 1, \ldots, k$, are assumed *unknown*; the conditional probability densities $f_i(x|\theta_i) = f(x|\theta_i, H_i)$ of an observation x, assuming it to come from class H_i, are assumed completely *known* (i.e. both the functional form and the parameter vectors θ_i are known). These assumptions may be appropriate when large training sets can be made available from each individual class, but there is little initial information regarding the "mix" of observations in the context under study.

Case B. The class probabilities, $\pi_1, \ldots, \pi_k$, are assumed *known*; the conditional $f_i(x|\theta_i) = f(x|\theta_i, H_i)$ are assumed to have known functional forms, depending on parameter vectors θ_i some, or all, of which are *unknown*. For example, in many contexts it may be appropriate to assume that underlying densities are Gaussian with unknown means, while the variances and the class probabilities are known.

Case C. The class probabilities $\pi_1, \ldots, \pi_k$ are assumed *unknown*; the conditional densities $f_i(x|\theta_i)$ are assumed to have a *known* functional form, depending on parameter vectors θ_i, some, or all, of which are *unknown*.

In all the cases, the problem is as follows. A sequence of (possibly vector-valued) observations, $x_1, \ldots, x_n, \ldots$ are received, one at a time, and each has to be classified as coming from one of a known number k of exclusive classes $H_1, \ldots, H_k$ before the next observation is received. Each decision is made on the basis of knowing all the previous observations, but without knowing whether previous classifications were correct or not. We assume that the x's are received at a high rate and that strict computational constraints are imposed. We thus limit ourselves to learning procedures whose demand for computational resources is small.

Defining $\psi = (\pi,\theta)$, where $\pi = (\pi_1, \ldots, \pi_k)$, $\theta = (\theta_1, \ldots, \theta_k)$, we assume that, conditional on ψ, the x_n are independent with probability density

$$f(x_n|\psi) = \sum_{i=1}^{k} \pi_i f_i(x_n|\theta_i), \tag{1}$$

(we shall assume throughout that the f's are such as to make this mixture identifiable (see Yakowitz, 1970)). For a sequence of observations, $x_1, \ldots, x_n$, it follows from (1) that

$$f(x_1, \ldots, x_n|\psi) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_i|\theta_j) \tag{2}$$

This is a sum of k^n products of component densities, each term in the summation having an interpretation as the probability of obtaining a certain partition of the observations among the classes.

The Bayesian algorithm for learning about ψ (or the components of interest) involves the specification of an a priori density for ψ, and the subsequent recursive computation of the posterior density $p(\psi|x_1, \ldots, x_n)$ using

$$p(\psi|x_1, \ldots, x_n) \propto f(x_n|\psi)p(\psi|x_1, \ldots, x_{n-1}) \tag{3}$$

Classification of x_n is based upon any specified loss structure, and for

$i = 1, \ldots, k$ the values of $p_r(x_n \epsilon H_i|x_1, \ldots, x_n)$ the probability that the n^{th} observation belongs to class H_i, given the observations $x_1, \ldots, x_n$. These probabilities are computed using

$$p_r(x_n \epsilon H_i|x_1, \ldots, x_n) \propto f_i(x_n|x_1, \ldots, x_{n-1}). \\ .p_r(x_n \epsilon H_i|x_1, \ldots, x_{n-1}) \tag{4}$$

here

$$f_i(x_n|x_1, \ldots, x_{n-1}) = \begin{cases} f_i(x_n|\theta_i), \text{ in Case A} \\ \int f_i(x_n|\theta_i)p(\theta_i|x_1, \ldots, x_{n-1})\, d\theta_i, \text{ in Case B,} \\ \int\int f_i(x_n|\theta_i)p(\theta_i,\pi|x_1, \ldots, x_{n-1})\, d\pi\, d\theta_i, \text{ in Case C,} \end{cases} \tag{5}$$

and

$$p_r(x_n \epsilon H_i|x_1, \ldots, x_{n-1}) = \begin{cases} \int \pi_i p(\pi_i|x_1, \ldots, x_{n-1}) d\pi_i, \text{ in case A} \\ \pi_i, \text{ in case B} \\ \int\int \pi_i p(\pi_i,\theta|x_1, \ldots, x_{n-1})\, d\theta\, d\pi_i, \text{ in case C} \end{cases} \tag{6}$$

It is obvious that due to the mixture form inherent in (1) and (2) there exist no reproducting (natural conjugate) densities for unsupervised Bayes learning. This results in an unavoidable increase in computer time and memory requirements, and leads to the solution being impractical in the case of signals arriving at a high rate, where speed of computation and small memory requirements are basic prerequisites for a solution. For this reason, the formal Bayes learning procedure (B) has been regarded as of little practical use. Among the ad hoc solutions proposed in its place, we note the Decision Directed approach, Recursive Moment Estimates and Learning with Probabilistic Teacher, all of which are discussed in the references given above.

As an alternative to these, we propose a Quasi-Bayes procedure which is both highly computationally efficient and retains the *flavour* of the formal

Bayes solution. Our discussion will be in terms of Cases A, B and C as above, but the approach can be extended to more general solutions. The statistical literature abounds with papers on the estimation of parameters of mixture distributions. The proposed methods (maximum likelihood estimators, moment generating function estimator, method of moments) demand considerable computational resources and thus will not be discussed here. (For references, see Quandt and Ramsey, (1978) and the ensuing discussion).

2. APPROXIMATE PROCEDURES FOR CASE A.

For convenience of notation, we shall write $\pi = (\pi_1,\ldots,\pi_k)$ for the unknown class probabilities, and $f_i(x_n)$ for the known densities. Prior knowledge about π is specified in the form of an a priori density $p(\pi)$.

If we denote by $p(\pi|X_n)$ the posterior density for π given $X_n = (x_1,\ldots,x_n)$, and by $p_i(\pi|X_n)$ the posterior density for π if it is also known that $x_n \in H_i$, then, by Bayes theorem,

$$p(\pi|X_n) = \sum_{i=1}^{k} w_i(X_n)\, p_i(\pi|X_n), \tag{7}$$

where

$$w_i(X_n) = p(x_n \in H_i|X_n) = \frac{f_i(x_n)\hat{\pi}_i(X_{n-1})}{\Sigma_{i=1}^{k} f_i(x_n)\hat{\pi}_i(X_{n-1})}; \tag{8}$$

and

$$\hat{\pi}_i(X_{n-1}) = \int \pi_i\, p(\pi|X_{n-1})d\pi. \tag{9}$$

We now consider the special case where $p(\pi)$ has the form of a Dirichlet density

$$p(\pi) = \frac{\Gamma(\alpha_1^{(0)}+\ldots+\alpha_k^{(0)})}{\Gamma(\alpha_1^{(0)})\ldots\Gamma(\alpha_k^{(0)})} \prod_{i=1}^{k} \pi^{\alpha_i^{(0)}-1} \tag{10}$$

which we denote by $D(\pi|\alpha_1^{(0)},\ldots,\alpha_k^{(0)})$, where $\Gamma(\cdot)$ is the standard gamma function. Such a form might arise, for example, following a multinomially distributed training sample whose correct classifications were known.

It follows from (7) and (10) that after observing x_1 we obtain

$$p(\pi|X_1) = \sum_{i=1}^{k} w_i(X_1)\, D(\pi|\alpha_1^{(0)} + \delta_{i1},\ldots,\alpha_k^{(0)} + \delta_{ik}), \tag{11}$$

where

$$w_i(X_1) = \frac{f_i(x_1)\alpha_i^{(0)}}{\sum_{i=1}^{k} f_i(x_1)\alpha_i^{(0)}} \qquad (i = 1,\ldots,k) \tag{12}$$

and

$$\delta_{ij} = 1 \text{ if } i = j,$$
$$= 0 \text{ otherwise.}$$

Many well-known approximate learning procedures for this problem can be seen as arising from approximations to (11) of the form

$$p(\pi|X_1) \approx D(\pi|\alpha_1^{(0)} + \hat{\delta}_{11},\ldots,\alpha_k^{(0)} + \hat{\delta}_{1k}), \tag{13}$$

where the $\hat{\delta}_{ij}$'s take values according to some specified method. Two approaches are suggested.

I. Averaging. The $\hat{\delta}_{ij}$'s are chosen such that the mean and variance of the approximating density (13) are identical to those of the mixture (11). A similar approach (though in a different context) is taken in Owen (1975); Athans, Whiting and Gruber (1977); Harrison and Stevens (1976).

II. Selection. Here one of the $\hat{\delta}_{ij}$ takes the value one and the others zero according to some decision rules. This approach is akin to the engineering concept of 'learning without a teacher', see Agrawala, (1973); Spragins (1966); Fralick (1967), where the unknown 'teacher', the $\hat{\delta}_{ij}$, is the missing label identifying the observation with its class. Particular examples are the Decision-Directed learning and the Probabilistic Teacher Scheme. A comparative study (in the context of jumps in linear systems) of several averaging and selection methods is given in Smith and Makov (1980).

(i) Decision-Directed Learning (DD)

According to the DD approach one of the $\hat{\delta}_{ij}$ is set equal to one and the others zero in such a way that using (4) and some specified loss function, this results in a minimum expected posterior loss. In other words, by setting $\hat{\delta}_{ij}$ to equal zero or one we regard our own (unconfirmed) classification as if it were true. For example, in Scudder (1965),* the $\hat{\delta}_{ij}$ was set to equal one if $w_i(X_1)$

* In context of Case B.

was maximized for $i = j$. The approach in effect assumed that the most likely H_i was, in fact, the true one (and thus zero one loss function assumed).

The DD scheme was further studied in Davisson and Schwartz (1970), where it was shown that the approach did not guarantee asymptotic unbiasedness and could also lead to problems of *runaways*. Runaway occurs when the scheme commits a sequence of errors resulting in a degradation of performance and consequent convergence to biased values. In Davisson and Schwartz (1970), Davisson (1970), the detection of signals in Gaussian noise was considered and bounds on the probability of runaway were provides using random walk theory. It was shown that except for very low signal to noise ratio, the probability of runaway of the class probabilities to the extreme values 0 and 1 was very small.

In Katopis and Schwartz (1972), a modified version of DD (MDD) was proposed in which a bias-removing transformation of the observations was introduced such that the convergence to the true value of the class probability π was ensured. Another modification was given in Schwartz and Katopis (1977). In Kazakos and Davisson (1979), in adittion to a bias-removing transformation, a specific gain function (in the DD recursion) was suggested that guaranteed fastest mean square error convergence of the estimates of the π_i's. All these modifications were shown to avoid the problems associated with the DD scheme, but at the expense of requiring numerical integration after each observation.

(ii) Learning with a Probabilistic Teacher (PT)

According to this scheme, proposed in Agrawala (1970), a randomized choice is made; $\hat{\delta}_{ij}$ being set equal to one with probability $w_j(X_1)$. In Silverman (1979), the theoretical properties of the PT for Case A were discussed; convergence was proved and asymptotic relative efficiency properties were examined.

The scheme which we propose is as follows:

Quasi-Bayes Learning (QB), see Makov and Smith (1977); Smith and Makov (1978); Makov and Smith (1976), replaces $\hat{\delta}_{ij}$ by $w_j(X_1)$, and so takes

$$p\,(\pi\,|X_1) \approx D\,(\pi\,|\alpha_1^{(1)},\ldots,\alpha_k^{(1)}), \tag{14}$$

where

$$\alpha_i^{(1)} = \alpha_i^{(0)} + w_i(X_1)\,(i=1,\ldots,k). \tag{15}$$

Subsequent updating proceeds in the same way, so that with $p\,(\pi\,|X_{n-1})$ having a Dirichlet form with parameters $\alpha_i^{(n-1)}$, it follows that $p\,(\pi\,|X_n)$ will be Dirichlet with parameters

$$\alpha_i^{(n)} = \alpha_i^{(n-1)} + w_i(X_n)\,(i=1,\ldots k), \tag{16}$$

where, corresponding to (12),

$$w_i(X_n) = \frac{f_i(x_n)\;\alpha_i^{(n-1)}}{\Sigma_{i=1}^k f_i(x_n)\;\alpha_i^{(n-1)}} \qquad (i=1,\ldots,k) \tag{17}$$

In the special case $k = 2$, the Quasi-Bayes procedure leads to recursive estimates of π_1 if the form

$$\hat{\pi}_1^{(n+1)} = \hat{\pi}_1^{(n)} - a_n(\hat{\pi}_1^{(n)} - w_1^{(n+1)}), \tag{18}$$

where

$$a_n^{-1} = (\alpha_1^{(0)} + \alpha_2^{(0)} + n + 1) \tag{19}$$

and

$$w_1^{(n+1)} = \frac{f_1(x_{n+1})\;\hat{\pi}_1^{(n)}}{f_1(x_{n+1})\hat{\pi}_1^{(n)} + f_2(x_{n+1})\hat{\pi}_2^{(n)}} \tag{20}$$

(18) is a typical QB recursion (for this case and others), which corresponds to a Robbins-Monro (Robbins and Monro, 1951) type of Stochastic Approximation. Using existing theorems in this field (e.g. Gladyshev, 1965, and many other) we were able to prove that the QB scheme converges to the true value of π in mean square and with probability one. Convergence properties were established for the case $k = 2$ in Makov and Smith (1977), Makov (1980), and for general k in Smith and Makov (1978). It was also shown in Makov and Smith (1976), that the QB scheme provides a better approximation to the Bayes solution than does the MDD. In Silverman (1979) the QB was proved to be more efficient than the PT.

In Kazakos (1977), a recursive estimation algorithm was provided which was based on the minimization of the Kullback-Leiber information number. The algorithm was shown to be consistent (for any k) and efficient (for $k=2$). In Makov (1980), it was shown that the QB scheme, (18) - (20), is a special case of the one of the discussed in Kazakos (1977).

In Fig. 1, we show the paths of successive estimates of π_1, π_2 for a three-class simulated example ($k=3$), where f_1, f_2, f_3 are circular bivariate Gaussian

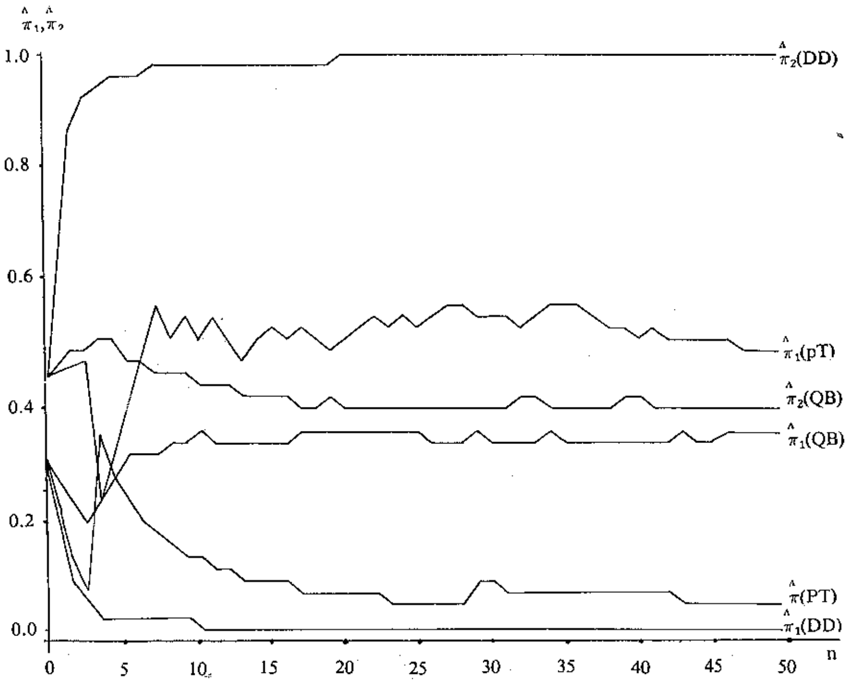


FIG. 1

distributions with all variances equal to one and means given by (-0.5,0), (0,0.5) and (0.5,0) respectively. Comparisons of the QB approach with the B solution have been made in Makov and Smith (1977), Smith and Makov (1978), and we have omitted calculation of B here. Comparison of QB with MDD was made in Makov and Smith (1976), where the latter was shown to be definitely inferior. Since the MDD would, in fact, require successive two-dimensional numerical integration for this example, it is also omitted. The results from the DD, PT and QB schemes are shown for the first 50 simulated observations, where π_1 and π_2 were both equal to 0.33. The estimates for QB were obtained using (9) and (16), which, from the well-known form of the mean of a Dirichlet distribution, implies that

$$\hat{\pi}_i (X_n) = \frac{\alpha_i^{(n)}}{\Sigma_{i=1}^k \alpha_i^{(n)}} \qquad (i=1,\ldots,k) \qquad (21)$$

The estimates for PT use successive randomizations as described above, or in Agrawala (1970); the estimates for DD follow the procedure as described above, or in Davisson and Schwartz (1970). The prior parameters used were $\alpha_1^{(0)} = 0.1$, $\alpha_2^{(0)} = 0.15$ and $\alpha_3^{(0)} = 0.1$, representing a very weak form of prior knowledge, implying prior means for π_1, π_2, π_3 of 0.286, 0.428 and 0.286, respectively.

In this and similar examples, where classification is made difficult because of the high overlap of the underlying distributions, the QB method shows marked superiority over the PT method, while the DD method performs very badly indeed. When the underlying distributions have only moderate overlap, there appears little to choose between QB and PT, whereas both are markedly superior to DD.

3. QUASI-BAYES PROCEDURES FOR CASE B

In order to illustrate our approach to problems which fall within the framework of Case B, we shall consider two special cases, both for the case $k=2$, and both involving known π_1, π_2 $(=1-\pi_1)$. The first is that of *Bipolar signal* detection, where $f_1(x|\theta_1)$ is a Gaussian density with unknown mean $\theta>0$, $f_2(x|\theta_2)$ is a Gaussian density with mean $-\theta$, and the variances are known and equal (to σ^2, say). The second is that of *Signal versus Noise* detection, where $f_1(x|\theta_1)$ is a Gaussian density unknown mean θ, $f_2(x|\theta_2) = f_2(x)$ is a Gaussian density with mean zero, and the variances are known and equal (to σ^2, say).

From the general results given in the introduction, it can be shown that if we take $p^o(\theta)$ to be normal with mean μ and variance τ^2, then after observing x_1 we have

$$p^{(1)}(\theta) = \Sigma_{i=1}^2 w_i^{(1)} N(\theta; \tau^{-2}\mu + \sigma^{-2}\hat{\delta}_{i1}x_1, \tau^{-2} + \sigma^{-2}|\hat{\delta}_{i1}|) \qquad (22)$$

where $w_i^{(1)} = p_r(x_1 \epsilon H_i | x_1)$ is derivable from (4), (5) and (6), $N(\theta; c,d)$ denotes that θ has Gaussian distribution with mean c/d, variance d^{-1}, and $\hat{\delta}_{ij} = 1$ or -1 according as $i = 1$ $(x_j \in H_1)$, or not, in the Bipolar signal case, $\hat{\delta}_{ij} = 1$ or 0 according as $i = 1$ $(x_j \in H_1)$, or not, in the Signal versus Noise case.

Our proposal is to replace $\hat{\delta}_{i1}$ by $E(\delta_{i1})$, which is equal to $2w_1^{(1)}-1$ in the Bipolar case, and equal to $w_1^{(1)}$ in the Signal versus Noise case, and to take $p^{(1)}(\theta) = N(\theta; \tau^{-2}\mu + \sigma^{-2} E(\delta_{i1})x_1, \tau^{-2} + \sigma^{-2}E(|\delta_{i1}|))$. Subsequent updating now takes place entirely within the Gaussian family, and we obtain

$$p^{(n)}(\theta) = N\left(\theta; \tau^{-2}\mu + \sigma^{-2}\,\Sigma_{j=1}^{n} E(\delta_{ij})x_j, \tau^{-2} + \sigma^{-2}\,\Sigma_{j=1}^{n} E(|\delta_{ij}|)\right). \tag{23}$$

The posterior means give a sequence of estimates of θ, and the following recursive relations are obtained:

For the Bipolar case

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} - \frac{\sigma^{-2}}{\tau^{-2} + (n+1)\sigma^{-2}}\{\hat{\theta}^{(n)} - (2w_1^{(n+1)}-1)x_{n+1}\} \tag{24}$$

For the Signal versus Noise case

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} - \frac{\sigma^{-2}}{\tau^{-2} + \sigma^{-2}\,\Sigma_{i=1}^{n+1} w_1^{(i)}}\{\hat{\theta}^{(n)} - x_{n+1}\, w_1^{(n+1)}\} \tag{25}$$

Various modifications can also be considered for large n, but these are not discussed here. In Smith and Makov (1981), the convergence properties of the Signal versus Noise scheme were discussed for the case where the w's are replaced by $w_j^{*(n)} = p_r(x_n \in H_i | \hat{\theta}^{(n-1)}, \pi)$. The resulting recursion was shown to converge to θ with probability one. In Titterington, 1976, a technique similar to the QB was applied in the context of medical diagnosis where unconfirmed cases (= unsupervised) were incorporated into data banks. The 'fractional updating' was used to estimate the means and covariance matrices of multivariate normal densities.

The performance of the DD scheme and its improved version have been studied in general in Patrick, Costello and Monds (1970); Young and Farjo (1972). At the present time, following the criticism of Agrawala (1970), made in Cooper (1975), there would appear to be no satisfactory account of the theoretical properties of the PT scheme for this case.

In Fig. 2, we show the paths of successive estimates for a simulated example of the Signal versus Noise problem. A comparison is given, for the first 50 observations, of the Decision Directed, Improved Decision Directed, Probabilistic Teacher and Quasi-Bayes methods. The underlying parameters were as follows: $\theta = 2.0$, $\sigma^2 = 4.0$, $\pi_1 = 0.5$, $\mu = 5.0$, $\tau^2 = 25.0$; the latter represent very vague prior knowledge about θ. Again the pattern shown by this example is typical. Both the Probabilistic Teacher and the Quasi-Bayes procedure perform better than the Decision Directed schemes.

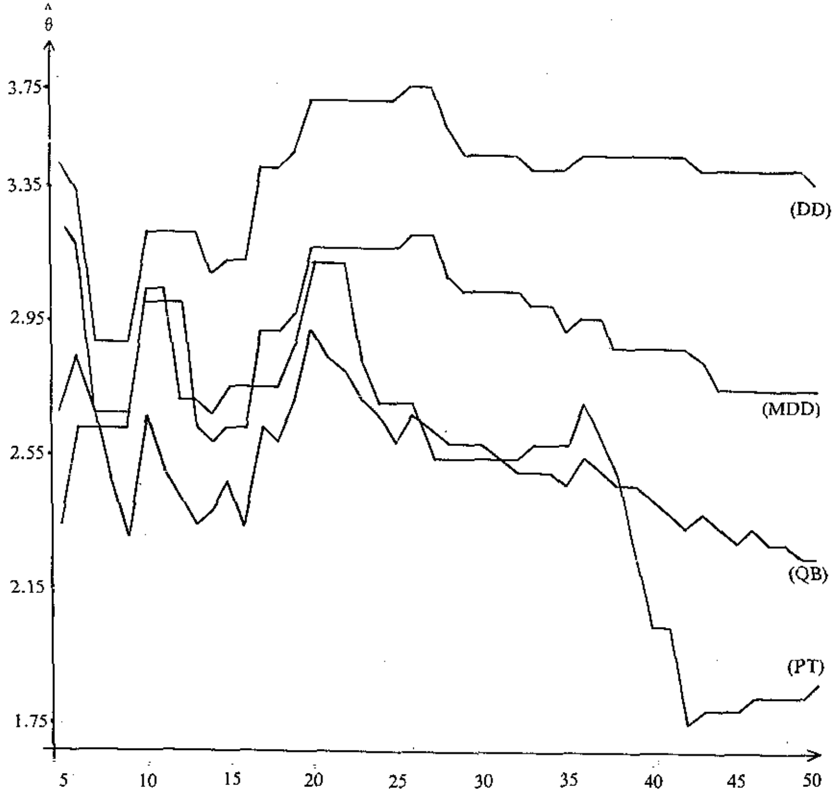


FIG. 2

4. QUASI-BAYES PROCEDURE FOR CASE C

Few results are available in this rather difficult case. In Young and Coraluppi (1970), stochastic estimation of a mixture of normal densities using an information criterion is discussed. In Katopis and Schwartz (1972); Schwartz and Katopis (1977), modified DD schemes proved to be consistent in a two-class decision problem where the mixture consisted of two normal densities, the mean of one of which was unknown (as well as the mixing parameter). In Makov (1980a), the QB scheme was attempted in a Kalman filter context in which an attempt was made to track a process when there was a non-zero probability that the observation contained nothing but pure noise. Simulations results showed that the QB scheme is by far more reliable than the PT or DD so long as the process is going through the contaminated environment. Work is in progress on the mathematical properties of the QB in

Case C. Preliminary results indicate that convergence may be guaranteed if certain restrictions are imposed on the parameter space. This will not be discussed here.

ACKNOWLEDGEMENT

I am grateful to Professor A.F.M. Smith for his useful comments on the manuscript.

REFERENCES

AGRAWALA, A.K. (1970). Learning with a probabilistic teacher. *IEEE Trans. Inform. Theory* **IT-16,** 373-379.

— (1973). Learning with various types of teachers. *Proc. 1st. Int. Joint Conf. Pattern Recognition,* 453-461.

ATHANS, M., WHITING, R.H. & GRUBER, M. (1977). A suboptimal estimation algorithm with probabilistic editing for false measurements with application to target tracking with wake phenomena. *IEEE Trans. on Automat. Contr.* **AC-22,** 372-384.

COOPER, D.B. (1975). On some convergence properties of 'learning with a probabilistic teacher' algorithms. *IEEE Trans. Inform. Theory* **IT-21,** 699-702.

DAVISSON, L.D. (1970). Convergence probability bounds for stochastic approximation. *IEEE Trans. Inform. Theory* **IT-16,** 680-685.

DAVISSON, L.D. and SCHWARTZ, S.C. (1970). Analysis of a decision directed receiver with unknown priors. *IEEE Trans. Inform. Theory* **IT-16,** 270-276.

FRALICK, S.C. (1967). Learning to recognize patterns without a teacher. *IEEE Trans. on Inform. Theory* **IT-13,** 57-64.

FU, K.S. (1968). *Sequential Methods in Pattern Recognition and Machine Learning.* New York: Academic Press.

GLADYSHEV, E.G. (1965). On stochastic approximation. *Theory of Prob. and its Appl.* **10,** 275-278.

HARRISON, P.J. and STEVENS, C.P. (1976). Bayesian forecasting. *J.R. Statist. Soc. B.* **38,** 205-247.

HO, Y.C. and AGRAWALA, A.K. (1968). On pattern classification algorithms; introduction and survey. *Proc. IEEE* **56,** 2102-2114.

KATOPIS, A. and SCHWARTZ, S.C. (1972). Decision directed learning using stochastic approximation. *Proc. Modelling and Simulation Conf.,*- 473-481.

KAZAKOS, D. (1977). Recursive estimation of prior probabilities using a mixture. *IEEE Trans. on Inform Theory* **IT-23,** 203-211.

KAZAKOS, D. and DAVISSON, L.D. (1979). An improved decision-directed detector. *IEEE Trans. on Inform. Theory.* Submitted to publication.

MAKOV, U.E. (1980). On the choice of gain functions in recursive estimation of prior probabilities. *IEEE Trans. on Inform. Theory.* **IT-26,** 497-498.

— (1980a). A quasi-Bayes approximation for unsupervised filters. *IEEE Trans. on Autom. Contr.* **AC-25,** 842-847.

MAKOV, U.E. and SMITH, A.F.M. (1976). Quasi Bayes procedures for unsupervised learning. *Proc. IEEE Conf. on Decision and Control.* 408-412. New York: IEEE Inc.

— (1977). A quasi-Bayes unsupervised learning procedure for priors. *IEEE Trans. Inform. Theory* **IT-23,** 761-764.

OWEN, J.R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *J. Amer. Statist. Assoc.* **70,** 351-356.

PATRICK, E.A. (1972). *Fundamentals of Pattern Recognition.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

PATRICK, E.A., COSTELLO, J.P. and MONDS, F.C. (1970). Decision directed estimation for a two class decision boundary. *IEEE Trans. on Computers* **C-19,** 197-205.

QUANDT, R.E. and RAMSEY, J.B. (1978). Estimating mixture of normal distributions and swithching regressions. *J. Amer. Statist. Assoc.* **73,** 730-738.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **25,** 400-407.

SCHWARTZ, S.C. and KATOPIS, A. (1977). Modified stochastic approximation to enhance unsupervised learning. *Proc. of the IEEE Conf. on Decision and Control.* 1067-1069.

SCUDDER, H.J. (1965). Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inform. Theory* **IT-11,** 363-371.

SILVERMAN, B.W. (1979). Some asymptotic properties of the probabilistic teacher. *IEEE Trans. on Inform. Theory.* **IT-26,** 296-249.

SMITH, A.F.M. and MAKOV, U.E. (1978). A quasi-Bayes sequential procedure for mixtures. *J. Roy. Statist. Soc. B* **40,** 106-112.

— (1980). Bayesian detection and estimation of jumps in linear systems. *Proceedings of the IMA Conference on the Analysis and Optimization of Stochastic Systems,* (O.L.R. Jacobs *et.al.,* eds), 333-346. New York: Academic Press.

— (1981). Approximation to Bayes learning procedures. *IEEE Trans. on Inform. Theory.* To appear.

SPRAGINS, J. (1966). Learning without a teacher. *IEEE Trans. on Inform. Theory* **IT-12,** 223-230.

TITTERINGTON, D.M. (1976). Updating a diagnostic system using unconfirmed cases. *Appl. Statist.* **25,** 238-347.

YAKOWITZ, S.J. (1970). Unsupervised learning and the identification of finite mixtures. *IEEE Trans. on Inform. Theory* **IT-16,** 330-338.

YOUNG, T.Y. and CORALUPPI, G. (1970). Stochastic estimation of a mixture of normal density functions using an information criterion. *IEEE Trans. on Inform. Theory* **IT-16,** 258-263.

YOUNG, T.Y. and FARJO, A.A. (1972). On decision directed estimation and stochastic approximation. *IEEE Trans. on Inform. Theory* **IT-18,** 671-673.

YOUNG, T.Y. and CALVERT, T.W. (1974). *Classification, Estimation and Pattern Recognition.* New York: American Elsevier.

Change-Point problems: approaches and applications

A.F.M. SMITH
University of Nottingham

SUMMARY

Problems of making inferences about abrupt changes in the mechanism underlying a sequence of observations are considered in both retrospective and on-line contexts. Among the topics considered are the Lindisfarne scribes problem; switching straight lines; manoeuvering targets, and shifts of level or slope in linear time series models. Summary analyses of data obtained in studies of schizophrenic and kidney transplant patients are presented.

Keywords: CHANGE-POINT; SWITCHING STRAIGHT LINES; BAYES FACTOR; KALMAN FILTER.

1. INTRODUCTION

In the simplest possible case, a sequence of random quantities $\tilde{y}_1,\ldots,\tilde{y}_n$ is said to have a change-point at r $(1 \leq r < n)$ if $\tilde{y}_1,\ldots,\tilde{y}_r$ and $\tilde{y}_{r+1},\ldots,\tilde{y}_n$ are exchangeable subsequences, but the combined sequence is not exchangeable. Assuming the usual mixture representation of exchangeable sequences, the most frequently used model of a change-point at r can be written in terms of densities as

$$p\,(y_1,\ldots,y_n|M_r) = \int\int \Pi_{i=1}^{r} p_1(y_i|\theta_1)\,\Pi_{i=r+1}^{n} p_2(y_i|\theta_2)p\,(\theta_1,\theta_2)d\theta_1 d\theta_2, \qquad (1)$$

where M_r denotes the model which assumes a change-point at r, and $p_1(y|\theta_1) \neq p_2(y|\theta_2)$, $p\,(\theta_1,\theta_2)$ have obvious interpretations. It is convenient to denote by M_0 the model which assumes the entire sequence exchangeable and defines

$$p\,(y_1,\ldots,y_n|M_0) = \int\int \Pi_{i=1}^{n} p_1(y_i|\theta_1)p\,(\theta_1,\theta_2)d\theta_1 d\theta_2. \qquad (2)$$

With such a formulation, inference about change-points, given $\tilde{y}_1 = y_1,\ldots,\tilde{y}_n = y_n$, reduces to consideration of the set of alternative models M_0, $M_1,\ldots,M_{n-1}$. These may be conveniently compared pairwise using Bayes factors - ratios of posterior to prior odds - so that, as is easily seen from Bayes theorem,

$$B_{ij} = \frac{p\,(y_1,\ldots,y_n\,|\,M_i)}{p\,(y_1,\ldots,y_n\,|\,M_j)}\;, \tag{3}$$

the required densities being obtained from (1) and (2). A detailed study of this approach for univariate sequences and a variety of standard parametric distributions is given in Smith (1975). In Section 2 of this paper, we shall outline the extension to more than one possible change-point and illustrate the approach by applying it to the Lindisfarne Scribes problem (Ross, 1950).

In the more general setting of changes in structure of a regression or time series model, the simple characterization in terms of exchangeable subsequences no longer applies, but, provided we specify the model, M_r, corresponding to a change at r, we can use (3) directly to compare alternative models. This approach will be presented for regression models in Section 3 and a possible extension to linear time series models will be outlined in Section 4. Also in Section 3, we shall comment briefly on special problems of interest that arise in the case of switching straight lines.

The analysis in Sections 2-4 concentrates on *retrospective* analysis. In Section 5, we shall consider an alternative linear model formulation, in terms of Kalman filters (Harrison and Stevens, 1976), that seems more suited to *on-line* detection of changes.

2. BINOMIAL DATA: THE LINDISFARNE SCRIBES PROBLEM

The Lindisfarne Scribes problem (Ross, 1950; Silvey, 1958) is of the type described at the beginning of Section 1, but admitting more than one possible change-point. A text divides into n sections, and it is assumed that only one scribe was involved in the writing of any one section, and that sections written by any one scribe are consecutive. We wish to infer how often, and where, changes of scribe occurred. The analysis is to be based on the frequency of occurrence of a certain word which has just two alternative forms. A version of some of the data, taken from Ross (1950), is set out in Table 1.

TABLE 1

Number of occurrences of present indicative 3rd. singular endings s and δ for different sections of Lindisfarne

Section

	1	2	3	4	5	6	7	8	9	10	11	12	13
s...	12	26	31	24	28	34	39	46	41	19	17	17	16
δ...	9	10	13	6	24	11	9	11	7	3	3	4	4
Total...	21	36	44	30	52	45	48	57	48	22	20	21	20

The assumption is made that a scribe is characterized by the propensity with which, when using the present indicative third person singular, he adopts one or other of the two variants. We thus arrive at an example of a change-point problem, with many possible changes, where it might be reasonable to assume underlying binomial distributions.

If $M\,(r_1,\ldots,r_K)$ is the model which assumes K changes of scribe, with change-points $r_1,\ldots,r_K$, then if $\theta_1,\ldots,\theta_{K+1}$ denote the propensities of the assumed $K+1$ scribes, and m_i,y_i, $i = 1,\ldots,n$, the numbers of word uses and δ-variant uses, respectively, in each of the n sections, we have

$$\begin{aligned} p\,[y_1,\ldots,y_n\,|\,M(r_1,\ldots,r_K)] &= \Pi_{i=1}^{n}\binom{m_i}{y_i}\ \mathrm{x} \\ &\int\cdots\int \Pi_{j=1}^{k+1}\theta_j^{s_j}\ (1-\theta_j)^{f_j}\,p\,(\theta_1,\ldots,\theta_{k+1})d\theta_1,\ldots,d\theta_{k+1}, \end{aligned} \tag{4}$$

where

$$\begin{aligned} s_j &= s(r_{j-1}+1,r_j) = y_{r_{j-1}+1}+\ldots+y_{r_j} \\ f_j &= f(r_{j-1}+1,r_j) = m_{r_{j-1}+1}+\ldots+\mathrm{m}_{r_j}-s(r_{j-1}+1,r_j). \end{aligned} \tag{5}$$

There are, of course, no general prescriptions for the choice of $p\,(\theta_1,\ldots,\theta_{K+1})$. In some change-point contexts, for example in reliability studies, one might expect monotonic relationships to hold (Smith, 1977), but for the purpose of this illustration we shall simply consider the (perhaps unreasonable) assignment of independent beta prior densities, so that

$$p\,(\theta_1,\ldots,\theta_{K+1}) = \Pi_{j=1}^{k+1}\theta_j^{\alpha_j-1}\,(1-\theta_j)^{\beta_j-1}\,/\,B\,(\alpha_j,\beta_j), \tag{6}$$

where $B(.,.)$ denotes the usual beta function. Substituting (6) in (4), the required integration is immediate, and it is easily seen, for example, that

$$B_{0,(r_1,\ldots,r_k)} = \frac{B(\alpha_1+s(1,n),\beta_1+f(1,n))}{B(\alpha_1,\beta_1)} \bigg/ \Pi_{j=1}^{K+1} \frac{B(\alpha_j+s(r_{j-1}+1,r_j),\beta_j+f(r_{j-1}+1,r_j))}{B(\alpha_j,\beta_j)} \tag{7}$$

For the particular choice $K = 1$, $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 1$, and converting (7) into posterior probabilities on M_0, $M_1,\ldots,M_{13}$ (taking prior probability ½ on M_0, 1/26 on the others) we obtain the results shown in Table 2.

TABLE 2

Posterior probabilities assuming at most one change-point

M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	M_{13}
-	-	-	-	.50	.41	.07	.02	-	-	-	-	-	-

If we go on to consider $K=2$, $\alpha_i=\beta_i=1$, $i=1,2,3$, and $Pr\,(K = 0) = Pr\,(K = 1) = Pr\,(K = 2) = 1/3$, with equal prior probabilities on all thirteen models, given that $K = 1$, and on all seventy-eight models, given that $K = 2$, we obtain the results shown in Tables 3 and 4.

TABLE 3

Posterior probabilities of up to two changes

no change	*one change*	*two changes*
0.00002	0.06856	0.93142

TABLE 4

Posterior probabilities of selected pairs of change-points

	5	6	7	8	9	10
1	.029	.029	-	-	-	-
2	.027	-	-	-	-	-
3	.037	.026	-	-	-	-
4	.287	.049	-	-	-	-
5	-	.042	.043	.048	.029	.029

The analysis so far would appear to indicate strongly that there was a change of scribe after section 4 and again after section 5. In fact, further analysis suggests that there is no strong evidence for further changes. As an example, we note that the Bayes factor for M (4,5) against M (4,5,6) is given, from (7), by

$$B_{0,(4,5,6)} / B_{0,(4,5)} = \frac{46.237}{282} \frac{\binom{45}{11}\binom{236}{41}}{\binom{281}{52}} \approx 3.29 \tag{8}$$

This is, of course, the same result as is obtained by taking sections 6-13 and testing for a change after section 6.

Finally, we note in passing that the calculations required in (7) can be greatly simplified by applying Stirling's approximation. For example, if we have $K = 1$ and define s_1, s_2, f_1, f_2 by (5), then (7) has the form

$$B_{0,r_1} = \frac{(s_1+f_1+1)\,(s_2+f_2+1)}{(n+1)} \, \frac{\binom{s_1+f_1}{s_1}\binom{s_2+f_2}{s_2}}{\binom{n}{s_1+s_2}} \tag{9}$$

which can be shown to be well approximated by

$$B_{0,r_1} = \left[\frac{n\,(s_1+f_1)\,(s_2+f_2)}{2\pi(s_1+s_2)\,(f_1+f_2)}\right]^{1/2} \exp\{-1/2\,\chi^2\}, \tag{10}$$

where

$$\chi^2 = \frac{n\,(s_1 f_2 - s_2 f_1)^2}{(s_1+f_1)(s_2+f_2)(s_1+s_2)(f_1+f_2)}, \tag{11}$$

the latter being the usual χ^2-statistic for testing the equality of the underlying propensities of two independent binomial samples. For (8), the approximation (10) gives 3.37.

At least in the case of a single change-point, the above approach has many points of contact with the Bayesian significance testing approaches of Jeffreys (1961) and Dickey and Lientz (1970).

3. CHANGE IN A REGRESSION RELATIONSHIP

We shall consider the problem of investigating the stability over time of the regression model

$$\tilde{y}_t = \mathbf{x}_t^T \beta^{(t)} + \tilde{\varepsilon}_t\,, \quad t = 1,\ldots,n, \tag{12}$$

where at time, t, $\tilde{y}_t$ is the observation on the dependent variable, $\mathbf{x}_t$ is the column vector of observations on p regressor variables (including, possibly, a constant), $\beta^{(t)}$ is the column vector of unknown regression coefficients and $\tilde{\varepsilon}_t$ is the error term, assumed normally distributed with mean zero and variance σ^2.

In this section, we shall work with independent, homoscedastic errors and non-stochastic regressor variables. In the next section, we shall show how to extend the approach to cover more general situations.

The regression structure defined by (12) will be said to have a change-point at r $(1 \leq r < n)$ if

$$\beta^{(1)} = \ldots = \beta^{(r)} = \beta,\ \beta^{(r+1)} = \ldots = \beta^{(n)} = \beta + \delta$$

with unknown $\delta \neq \mathbf{0}$. We shall denote this model by M_r. The model of no change, $\delta = \mathbf{0}$, will be denoted by M_0.

If we adopt the notation,

$$\tilde{\mathbf{y}}_r^T = (\tilde{y}_1,\ldots,\tilde{y}_r), \quad \tilde{\mathbf{y}}_{(n-r)}^T = (\tilde{y}_{r+1},\ldots,\tilde{y}_n),$$
$$\mathbf{X}_r^T = (\mathbf{x}_1,\ldots,\mathbf{x}_r), \quad \mathbf{X}_{(n-r)}^T = (\mathbf{x}_{r+1},\ldots,\mathbf{x}_n),$$

we see that model M_r $(1 \leq r < n)$ can be written in the form

$$\tilde{\mathbf{y}} \sim N(\mathbf{A}_r\theta, \sigma^2\mathbf{I}_n), \tag{13}$$

where $\tilde{\mathbf{y}} = \tilde{\mathbf{y}}_n$, $\mathbf{I}_n$ is the n x n identity matrix and

$$\mathbf{A}_r = \begin{bmatrix} \mathbf{X}_r & \mathbf{0} \\ \mathbf{X}_{(n-r)} & \mathbf{X}_{(n-r)} \end{bmatrix}, \ \theta = \begin{bmatrix} \beta \\ \delta \end{bmatrix} \tag{14}$$

In the case of M_0, (13) still holds, but with $\mathbf{A}_0 = \mathbf{X}_n$, $\theta = \beta$.

Again, inference about the change-point (is there one? and, if so, where?) reduces to consideration of the possible models M_r. To calculate (3) in this case, we require

$$p\,(\mathbf{y} \mid M_r) = \int\cdots\int p\,(\mathbf{y} \mid \mathbf{A}_r,\, \theta,\, \sigma)\, p\,(\theta,\, \sigma \mid \mathbf{A}_r)\; d\theta d\sigma, \tag{15}$$

and thus need to specify $p\,(\theta,\sigma \mid \mathbf{A}_r)$. This specification, and its relation to the whole question of significance tests and choice procedures among alternative linear models has been discussed at some length in the literature. A recent discussion is given by Smith and Spiegelhalter (1980).

In this paper, we shall examine the consequences of the specification,

$$p\,(\theta,\sigma \mid \mathbf{A}_r) = p\,(\theta \mid \mathbf{A}_r,\sigma)\, p\,(\sigma), \tag{16}$$

where $p\,(\sigma) \propto \sigma^{-1}$, and $p\,(\theta \mid \mathbf{A}_r,\sigma)$ corresponds, for $1 \leq r < n$, to a normal distribution with mean θ_0 and covariance matrix $\sigma^2\mathbf{V}_0$, where

$$\mathbf{V}_0 = \begin{bmatrix} \mathbf{V}_{0\beta} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{0\delta} \end{bmatrix}, \ \theta_0 = \begin{bmatrix} \beta_0 \\ \delta_0 \end{bmatrix}; \tag{17}$$

in the case of M_0, we simply have $\mathbf{V}_0 = \mathbf{V}_{0\beta}$, $\theta_0 = \beta_0$.

With this prior specification, it is easily verified that, performing the integration with respect to θ in (15), $p\,(\mathbf{y} \mid M_r,\sigma)$ is equal to

$$(2\pi\sigma^2)^{-n/2}\ |\mathbf{V}_0|^{-1/2}\, |\mathbf{V}_0^{-1} + \mathbf{A}_r^T A_r|^{-1/2}$$
$$\exp\{-(1/2\sigma^2)[R_r + (\hat{\theta}-\theta_0)^T(\mathbf{V}_0 + (\mathbf{A}_r^T\mathbf{A}_r)^{-1})^{-1}\,(\hat{\theta}-\theta_0)]\}, \tag{18}$$

where $\hat{\theta}$ denotes the usual least-squares estimate of θ, and R_r the corresponding residual sum of squares.

If $\mathbf{V}_0^{-1}$ may be considered *small* in relation to $\mathbf{A}_r^T\mathbf{A}_r$, (18) may be simplified somewhat to give

$$p\ (\mathbf{y}\,|\,M_r,\sigma) \approx (2\pi\sigma^2)^{-n/2}|\mathbf{V}_{0\theta}|^{-1/2}|\mathbf{V}_{0\delta}|^{-1/2}|\mathbf{A}_r^T\mathbf{A}_r|^{-1/2}\exp\{-R_r/2\sigma^2\}, \qquad (19)$$

and

$$p\ (\mathbf{y}\,|\,M_0,\sigma) \approx (2\pi\sigma^2)^{-n/2}|\mathbf{V}_{0\theta}|^{-1/2}|\mathbf{A}_0^T\mathbf{A}_0|^{-1/2}\exp\ \{-R_0/2\sigma^2\}. \qquad (20)$$

Noting that $|\mathbf{A}_r^T\mathbf{A}_r| = |\mathbf{X}_r^T\mathbf{X}_r|\ |\mathbf{X}_{(n-r)}^T\mathbf{X}_{(n-r)}|$, the Bayes factor for M_0 against M_r, conditioned on known σ, is seen to be,

$$B_{0r}^{(\sigma)} = \left(\frac{|\mathbf{V}_{0\delta}|\,|\mathbf{X}_r^T\mathbf{X}_r|\,|\mathbf{X}_{(n-r)}^T\mathbf{X}_{(n-r)}|}{|\mathbf{X}_n^T\mathbf{X}_n|}\right)^{1/2}\exp\{-(1/2\sigma^2)(R_0\text{-}R_r)\}. \qquad (21)$$

Integrating (19) and (20) with respect to the assumed form for $p\ (\sigma)$, we obtain the unconditional Bayes factor

$$B_{0r} = \left(\frac{|\mathbf{V}_{0\delta}|\,|\mathbf{X}_r^T\mathbf{X}_r|\,|\mathbf{X}_{(n-r)}^T\mathbf{X}_{(n-r)}|}{|\mathbf{X}_n^T\mathbf{X}_n|}\right)^{1/2}\left(1 + \frac{p}{(n\text{-}2p)}F_r\right)^{-n/2}; \qquad (22)$$

where $F_r = [(R_0\text{-}R_r)/p\ ]/[R_r/(n\text{-}2p\ )]$ is the usual F-statistic for testing M_0 versus M_r.

In the special case of a univariate normal distribution with prior variance $\lambda\sigma^2$ for δ, the Bayes factor (22) reduces to

$$B_{0r} = \left(\frac{\lambda r(n\text{-}r)}{n}\right)^{1/2}\left(1 + \frac{t_r^2}{n\text{-}2}\right)^{-n/2}; \qquad (23)$$

where t_r is the two-sample t-test statistic corresponding to the samples $y_1,\ldots,y_r$ and $y_{r+1},\ldots,y_n$. The form (23) is similar to that derived for the two-sample problem by Jeffreys (1961, see comments following (13) of Section 5.41).

Application of (22) to the case of switching straight-lines has been made by Smith and Cook (1980). In this case, if δ_1, δ_2 are the components of δ representing possible changes in intercept and slope, respectively, then the parameter of interest is often $\gamma = -\delta_1/\delta_2$, the intersection point between the two straight-lines. Two cases are possible, according as a change at r necessarily implies $x_r \leq \gamma < x_{r+1}$, or not, where $x_1 < x_2 < \ldots < x_n$ denote the (time) ordered x-values. In the unconstrained case, we need to calculate

$$p(\gamma\,|\,\mathbf{y}) = \Sigma_r p(\gamma\,|\,r,\mathbf{y})p(r\,|\,\mathbf{y})\,,$$

the latter term being calculated using an appropriate transformation. In the constrained case, denoted by c, say, we require

$$p\ (\gamma\,|\,c,\mathbf{y}) = \Sigma_r\,p\ (c\,|\,\gamma,r,\mathbf{y})p\ (\gamma,r\,|\,\mathbf{y})/p\ (c\,|\,\mathbf{y}), \qquad (24)$$

where

$$P\,(c\,|\,\gamma,r,\mathbf{y}) = \begin{cases} 1 & \\ & \text{if} \\ 0 & \end{cases} \begin{matrix} \gamma \in (x_r,x_{r+1}) \\ \\ \gamma \notin (x_r,x_{r+1}) \end{matrix}$$

and

$$p\ (c\,|\,\mathbf{y}) = \Sigma_r\ \{\textstyle\int p\ (c\,|\,\gamma,r,\mathbf{y})p\ (\gamma\,|\,r,\mathbf{y})\}p\ (r\,|\,\mathbf{y}) = \Sigma_r \int_{x_r}^{x_{r+1}} p\ (\gamma,r\,|\,\mathbf{y})d\gamma.$$

Similarly, we can obtain

$$p\ (r\,|\,c,\mathbf{y}) = \int_{x_r}^{x_{r+1}} p\ (\gamma,r\,|\,\mathbf{y})d\gamma/\Sigma_r\int_{x_r}^{x_{r+1}} p\ (\gamma,r\,|\,\mathbf{y})d\gamma. \qquad (25)$$

These results were applied in Smith and Cook (1980) to data from kidney transplant patients, with the object of inferring the time of rejection of transplanted kidneys. It is thought that the constrained switching straight-line model provides a good model of the behaviour of reciprocal body-weight corrected serum-creatinine over the days following a transplant. Table 5 summarizes the data from a particular patient and the result from (25) when large prior variances are attached to the straight-line parameters and all change points are equally likely. The posterior density for γ given by (24) is symmetric and sharply peaked, with a mode at 4.15 and an approximate 95% credible interval is given by (3.71,4.59).

TABLE 5

Renal transplant data and posterior probabilities for r

r	1	2	3	4	5	6	7	8
y_r	48.4	58.3	62.3	73.1	68.3	55.3	49.1	43.9
$p(r\|c,y)$	—	.012	.316	.657	.012	.003	—	—

Retrospective studies of this kind are proving valuable in identifying patterns in the time to rejection of transplants and seem to have removed a great deal of the arbitrariness arising from doctors' attempts to "eyeball" the data. On-line analysis of this kind of data will be considered in Section 5.

Related material on switching straight lines can be found in Ferreira (1975).

4. SHIFT OF LEVEL IN AN ARMA PROCESS

In order to illustrate a reasonably straightforward extension of the approach of Section 3 to cover more general linear time series models, we shall consider the problem of investigating a shift in level of an ARMA (1,1) process. The material in this and the previous section is a direct development of some preliminary ideas given in Smith (1976).

We shall consider the following representation of a stationary ARMA (1,1) process with unknown mean level λ, and a shift in mean level of unknown magnitude δ occurring between the r^{th} and $(r+1)^{th}$ observations, where r is unknown. Let

$$
\begin{aligned}
\tilde{z}_1 &= \lambda + \tilde{\varepsilon}_1 \\
\tilde{z}_t &= \lambda + \tilde{\varepsilon}_t + (\varrho-\phi)\,\Sigma_{s=1}^{t-1}\varrho^{s-1}\tilde{\varepsilon}_{t-s}, \quad t = 2,\ldots,r, \\
\tilde{z}_t &= \lambda + \delta + \tilde{\varepsilon}_t + (\varrho-\phi)\,\Sigma_{s=1}^{t-1}\varrho^{s-1}\tilde{\varepsilon}_{t-s}, \quad t = r+1,\ldots,n,
\end{aligned}
\tag{26}
$$

with λ, δ, ϱ, ϕ and r unknown, $|\varrho|<1$, $|\phi| < 1$ and $\tilde{\varepsilon}_t$ independently and normally distributed with mean 0 and variance σ^2, with σ^2 unknown.

In order to utilize the development of Section 3, we make, conditional on ϕ and ϱ, the transformations

$$\tilde{y}_1 = \tilde{z}_1, \quad \tilde{y}_t = \tilde{z}_t - (\varrho-\phi)\,\Sigma_{s=1}^{t-1}\,\phi^{s-1}\,\tilde{z}_{t-s}, \quad t = 2,\ldots,n. \tag{27}$$

It is then easily seen that the vector $\mathbf{y}^T = (y_1,\ldots,y_n)$ satisfies (13), where, for $r \neq 0$,

$$\mathbf{A}_r^T = \mathbf{A}_r^T(\varrho,\phi) = \begin{pmatrix} a_1\ldots a_r & a_{r+1}\ldots a_n \\ 0\ldots 0 & a_1\ldots a_{n-r} \end{pmatrix}, \theta = \begin{pmatrix} \lambda \\ \delta \end{pmatrix} \quad ; \tag{28}$$

with $a_1 = 1$, $a_t = 1 - (\varrho-\phi)\,\Sigma_{s=1}^{t-1}\phi^{s-1}$, $t = 2,\ldots,n$. If $r = 0$, the "no-change" model, then $\theta = \lambda$ and $\mathbf{A}_0^T$ consists of just the first row of the matrix in (28).

By considering appropriate limits corresponding to $\varrho = 1$, $\phi = 0$ and $\varrho = 0$, respectively, the above framework can be used to study the special cases of IMA(1), AR(1) and MA(1) models. Related material can be found in Box and Tiao (1965) and Smith (1976).

Noting that the Jacobian of the transformation from $\mathbf{z}$ to $\mathbf{y}$ is unity, and denoting by $p\,(M_r,\varrho,\phi)$ a prior specification for M_r, ϱ and ϕ, we see from the results of Section 3 that $p\,(M_r,\varrho,\phi\,|\,\mathbf{z})$ is proportional to

$$(V_\lambda V_\delta)^{1/2}\,|\mathbf{A}_r^T(\varrho,\phi)\mathbf{A}_r(\varrho,\phi)|^{-1/2}\,(R_r(\varrho,\phi))^{-n/2}\,p\,(M_r,\varrho,\phi), \tag{29}$$

where $V_\lambda\sigma^2$, $V_\delta\sigma^2$ are the prior variances (conditional on σ) for λ and δ, and $R_r(\varrho,\phi)$ denotes the residual sum of squares from a least squares fit of M_r, given ϱ and ϕ.

The matrix whose determinant is to be evaluated in (29) has elements $a_1^2 + \ldots + a_n^2$ and $a_1^2 + \ldots + a_{n-r}^2$ on the diagonal, and $a_1 a_{r+1} + \ldots + a_{n-r}\,a_n$ as off-diagonal entries. The determinant and inverse are thus easily calculated.

Assignment of the prior probabilities for M_r, ϱ and ϕ depends, of course, on the situation under study. In any case, it seems that perfectly adequate results can be obtained by the crude form of numerical integration resulting from a suitable discretization of the ranges of ϱ and ϕ, so that calculation of marginal posterior probabilities are simply obtained from (29) by summation over the remaining variables. Inferences about λ, δ or σ^2, or predictive distributions for future observations, are obtained by forming weighted averages, with weights given by $p\,(M_r|\,\mathbf{z})$, of the standard results obtained by conditioning on a particular model M_r.

The procedure outlined above has been applied to a series of daily measurements of the time (in seconds) taken by an individual performing a certain psychological test repeated on 33 successive days. The data are presented in Table 6.

TABLE 6

Psychological test data

4.09	3.52	3.72	4.43	3.97	3.85	3.65	3.31	3.55	3.47	4.32
3.77	3.77	3.90	4.05	3.97	3.64	4.28	3.83	3.91	3.44	3.77
3.40	3.29	3.21	2.95	3.13	2.97	3.25	2.95	4.18	3.65	3.03

The individual has already passed through a "learning" phase on this test and it is believed that the observations would follow a stationary process, except that during this period of 33 days there has been a switch in background treatment regime. It is thought that this could have the effect of causing a sudden shift in performance level. The data were originally given to us with no information about where the change in treatment regime occurred. In fact, the change occurred between the 20th and 21st days.

Preliminary exploration of similar, unchanged, sequences of observations suggested that either an ARMA(1,1) or an AR(1) model might be suitable, and two corresponding analyses of the data were made. The first analysis assumed an ARMA (1,1) model with uniform priors over the range of r, the range of ϱ from -0.95 to 0.95 and ϕ between 0.000 and 0.95, the latter two in steps of 0.05. The second analysis considered an AR(1) model with a uniform prior for ϱ over the range -0.95 to 0.95. A summary of the results obtained are given in Table 7. No specification for V_λ is required, and V_δ is taken equal to 3.

TABLE 7

Summary inferences from the psychological test data

Posterior		
Summary	ARMA(1,1)	AR(1)
mean	21	21
r mode	22	22
median	22	22
ϱ mode	-	0.17
ϱ joint	0.17	-
ϕ mode	0.00	-
λ mean	3.83	3.83
δ mean	-0.53	-0.52

5. ON-LINE DETECTION OF CHANGE

Detailed description of the use of a set of alternative Kalman filter models has been given by Harrison and Stevens (1976) in the context of adaptive Bayesian forecasting procedures, and by Smith and Makov (1980) in the context of jump detection and estimation in linear systems, as required, for example, in the tracking of manoeuvering targets.

A general formulation allowing for sudden perturbations in either or both of the system and observation equations is given by representing model M_i, at time t, by

$$\theta_t = \mathbf{G}_{t-1}\theta_{t-1} + \mathbf{B}^{(i)}(\triangle\theta)_t + \mathbf{H}_{t-1}(\delta\theta)_t, \tag{30}$$

$$\mathbf{y}_t = \mathbf{F}_t\theta_t + \mathbf{C}^{(i)}(\triangle\mathbf{y})_t + (\delta\mathbf{y})_t, \tag{31}$$

where $(\triangle\theta)_t$, $(\triangle\mathbf{y})_t$ represent possible abrupt changes in either the system or the measurement at time t, $\mathbf{B}^{(i)}$, $\mathbf{C}^{(i)}$ define the specific nature of these changes according to model M_i, and $(\delta\theta)_t$, $(\delta\mathbf{y})_t$ are the usual Gaussian "noise" inputs to the system and measurement equations. The matrices $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$ define the general characteristics of the system.

In the case of manoeuvering targets, θ_t represents position and velocity components in some chosen frame of reference and $\mathbf{y}_t$ usually consists of observed position components. If $(\triangle\theta)_t$ consists of a finite set of plausible manoeuvres available at time t, models corresponding to particular choices of manoeuvre are defined by appropriate choices of $\mathbf{B}^{(i)}$ (assuming here that $\mathbf{C}^{(i)} = \mathbf{0}$).

In the case of the very useful univariate Linear Growth model (Harrison and Stevens, 1976, 3.4), the case of no abrupt change is modelled by

$$y_t = \mu_t + (\delta y)_t$$

$$\mu_t = \mu_{t-1} + \beta_t + (\delta\mu)_t$$

$$\beta_t = \beta_{t-1} + (\delta\beta)_t,$$

which can be represented in terms of (30) and (31) by

$$\begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix} = \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mu_{t-1}\\ \beta_{t-1}\end{pmatrix} + \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}(\delta\mu)_t\\ (\delta\beta)_t\end{pmatrix}$$

$$y_t = (1 \quad 0)\begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix} + (\delta y)_t$$

with $\mathbf{B}^{(i)} = \mathbf{C}^{(i)} = \mathbf{0}$. If we define this *no change* model to be M_0, and define M_1, M_2, M_3, by

$$\mathbf{B}^{(1)} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} ; \quad \mathbf{B}^{(2)} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} ; \quad \mathbf{B}^{(3)} = \mathbf{0},$$

$$\mathbf{C}^{(1)} = \quad 0 \quad ; \quad \mathbf{C}^{(2)} = \quad 0 \quad , \quad \mathbf{C}^{(3)} = 1,$$

we can represent "sudden change in level", "sudden change in slope" and "outlying observation", respectively, as well as "no change".

The recursive updating of the system, given a choice of M_i at time t, proceeds straightforwardly using the standard Kalman filter equations. Posterior weights on the individual models are also easily obtained using the appropriate modification of (3). In fact, of course, there is the problem of expanding mixture forms of posterior distribution, resulting from the unsupervised learning context, and practical use of this approach requires approximation of this mixture, at each stage, by a simple Gaussian distribution having the same mean and covariance structure as the mixture: see Harrison and Stevens (1976, 5.4) and Smith and Makov (1980) for further details.

This Linear Growth model, with the four model variants outlined above, has been used for on-line monitoring of kidney transplant patients, given data of the type shown in Table 5. For many patients, the series is considerably longer than the one shown, but we shall illustrate our procedure with this small data set. Table 8 shows, for each of the first six observations, the probability that it came from the situation modelled by M_0, M_1, M_2 or M_3. In addition, the table shows the same probabilities one-step back and two-steps back: thus, for example, $p\,(\tilde{y}_5 \epsilon M_2 | y_1,\ldots,y_7) = 0{\cdot}68$. By studying the changing pattern of these probabilities, the doctor can, hopefully, react to genuine changes fairly quickly, whilst avoiding over-hasty reactions to outlying measurements. Of course, the system depends on a number of prior inputs regarding reasonable variance levels and other features. These are assessed from knowledge of serum-creatinine measurement procedures and other background physiological information. Full details of this and other case studies will be reported elsewhere. The prior probabilities set on the four models for the first observation in this case were: 0.96, 0.01, 0.01, 0.02.

The results indicate that at observation 6 we suspect a slope change has occurred at observation 5. When we reach observation 7, we are fairly convinced that a slope change has occurred and that the patient is now in a new *steady state*. Posterior means of the slope parameter are positive up to and including observation 5 and then they suddenly switch to negative values, reinforcing the message of Table 8.

TABLE 8

On-line probabilities of M_0, M_1, M_2, M_3

Observation

	1				2				3			
	M_0	M_1	M_2	M_3	M_0	M_1	M_2	M_3	M_0	M_1	M_2	M_3
0-back	.99	-	-	-	.99	-	-	-	.99	-	-	-
1-back	.99	-	-	-	.99	-	-	-	.99	-	-	-
2-back	.99	-	-	-	.99	-	-	-	.99	-	-	-

	4				5				6			
	M_0	M_1	M_2	M_3	M_0	M_1	M_2	M_3	M_0	M_1	M_2	M_3
0-back	.99	-	-	-	.96	-	.01	-	.64	.09	.09	.17
1-back	.98	-	-	-	.56	.02	.41	.01	.84	.05	.10	-
2-back	.98	-	-	-	.29	.02	.68	-	.85	.04	.10	-

ACKNOWLEDGEMENTS

Study of kidney transplantation is in collaboration with Dr. M. Knapp of the Renal Unit, City Hospital, Nottingham. Study of schizophrenic test data is in collaboration with Dr. A. Elithorn of the Royal Free Hospital, London. The material of Section 4 has been developed in a Ph.D. thesis by N.B. Booth of the North London Polytechnic. The computations reported in Section 5 were carried out by M. West of Nottingham University.

REFERENCES

BOX, G.E.P. and TIAO, G.C. (1965). A change in level of a non-stationary time series. *Biometrika* **52**, 181-92.

DICKEY, J.M. and LIENTZ, B.P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann. Math. Statist.* **41**, 214-26.

FERREIRA, P.E. (1975). A Bayesian analysis of a switching regression model. *J. Amer. Statist. Assoc.* **70**, 370-74

HARRISON, P.J. and STEVENS, C.F. (1976). Bayesian Forecasting (with discussion). *J. Roy. Statist. Soc. B* **38**, 205-47.

JEFFREYS, H. (1961). *Theory of Probability*. Oxford: University Press.

ROSS, A.S.C. (1950). Philological Probability Problems. *J. Roy. Statist. Soc. B* **12**, 19-40.

SILVEY, S.D. (1958). The Lindisfarne Scribes Problem. *J. Roy. Statist. Soc. B* **20**, 93-101.

SMITH, A.F.M. (1975). A Bayesian approach to inference about a change point in a sequence of random variables. *Biometrika* **63**, 407-16.

— (1976). A Bayesian analysis of some time-varying models; In *Recent Developments in Statistics* (ed. Barra *et al*). 257-67. Amsterdam: North-Holland.

— (1977). A Bayesian note on reliability growth during a development testing program. *IEEE Trans. on Reliability* **R-26**, 346-47.

SMITH, A.F.M. and COOK, D.G. (1980). Switching straight lines: a Bayesian analysis of some renal transplant data. *Appl. Statist.* **29**, 180-89.

SMITH, A.F.M. and MAKOV, U.E. (1980). Bayesian detection and estimation of jumps in linear systems. *Proceedings of the IMA Conference on "The Analysis and Optimization of Stochastic Systems"*, (O.R.L., Jacobs *et. al.* ed.) 333-346. London: Academic Press.

SMITH, A.F.M. and SPIEGELHALTER, D.J. (1980) Bayes factors and choice criteria for linear models. *J. Roy. Statistic. Soc* **B. 42**, 213-20.

Discontinuity, decision and conflict

P.J. HARRISON and J.Q. SMITH

Warwick University

SUMMARY

The motivation for this paper arises out of the authors experiences in modelling real decision makers where the decisions show not only a continuous response to a continuously changing environment but also sudden or discontinuous changes. The theoretical basis involves a parametric characterisation of the environment, a decision makers perception of it in terms of a twice differentiable Distribution Function and a bounded Loss Function. Under a specified minimizing dynamic, the resultant Expected Loss Function satisfies the conditions for a potential function and Thoms Catastrophe Classification Theorem may be used to assess the singularity points and the thresholds at which jump decisions are taken. The paper describes the theory, summarises some results on unimodal distributions illustrated by jump decisions and population polarisation. Mixture distributions are then examined and the E* models defined. These are then briefly illustrated by reference to models which have been constructed in relation to Prison Riots, Agricultural and Economic modelling.

Keywords: BAYES; CATASTROPHE THEORY; CONFLICT MODELS; CONJUGATE LOSS FUNCTIONS; DECISION THEORY; DISCONTINUITY; E* MODELS; EXPECTED LOSS; ECONOMIC MODELS; FORECASTING; MIXTURES OF NORMALS; MULTIMODAL DISTRIBUTIONS; PRISON DISTURBANCES; POLARISATION; POTENTIAL FUNCTION; SMITH'S THEOREM; STEP LOSS FUNCTIONS; SEQUENTIAL INFORMATION.

1. INTRODUCTION

This paper is concerned with one way of viewing conflict. The stimulus arises from practical contexts met in macro modelling in the areas of agriculture, company cash flow, human relations, and managerial decision making. Briefly the approach adopted in practical modelling has been to consider the key decision makers in relation to their environment. The emphasis has been not only on the environmental response to decisions but on the decision makers responses to the environment. Many modellers ignore the latter which is so often vital in deriving forecasts and making the best use of them. Furthermore, it is not a thing encouraged by decision makers who tend to over react with statements that people are being treated as machines and free-will is being challenged. The latter is not so. Rather a decision maker is viewed perhaps as a farmer who is limited by his environment, but understands these limitations and takes planting decisions to act in harmony with the seasons. Unfortunately so often in the sphere of expedient socio-economic decision making, lags and feedbacks are ignored. Thus the usual decision process becomes totally out of phase with the requirements of the situation and the pronounced *medium* term 'expectations' of the decision makers, rather like a farmer planting at the worst possible time of year. Hence the decisions are eventually contradicted by hard fact giving rise to sudden policy reversals and outbreaks of conflict.

Theoretical models involving both continuous and discontinuous responses to continuous environmental change formed the basis of a Warwick Ph. D. for Jim Q. Smith (1978) and a resultant paper Smith, Harrison and Zeeman (1980). The approach involved a Bayesian decision theoretic formulation and utilises recent work in Catastrophe Theory (Thom, 1972). The multi-process models (Harrison and Stevens, 1971 and 1976) naturally lead to multimodal beliefs and via sensible utilities to multimodal expected loss functions on which Bayesian decision *makers base their decisions.* Similarly these functions can result from multimodal utilities and it is the presence of a multimodal expected loss function which here signals conflict or potential conflict. It is not pretended that this theoretical approach is the definitive way of modelling conflict. Rather it is one way of viewing situations and as such might catalyse and clarify insights into practical problems. This is particularly true where we consider an expected loss function which is a *potential function and can* be studied using the geometry of elementary catastrophes. Poston and Woodcock (1974), Zeeman (1977) and Poston and Stewart (1978).

This paper is very dependent upon Smith, Harrison and Zeeman (1980) and of course Smith (1978) for elaboration of some of the results and for background on the relation of catastrophes to decision sets. Section 2 of the paper describes the well known derivation of the Expected Loss Function and makes a number of definitions. Section 3 reviews some results on unimodal beliefs which, when combined with symmetric loss functions, give rise to conflicting decisions via multi-modal expected losses. An example involving log-Normal beliefs and a double step loss function is given. A theorem is then stated which forms the basis for an example in which a Pareto belief distribution combines with a double step loss function to give a two point Bayesian decision set. This might be used to model the polarisation of opinion over a population showing no such polarisation in beliefs or loss function.

Section 4 discusses some contexts in which multimodal phenomena can arise and looks at duality. Section 5 is a direct excerpt from Smith, Harrison and Zeeman (1980) stating Smith's Theorem concerning bi-modality.

The following sections all concentrate on beliefs and loss functions which are mixtures of Normals and mixtures of their conjugate loss functions (*Lindley, 1976*). These are introduced in section 6 where the $E^*_n[\delta]$ model is defined and the particular case of a $E^*_1[\delta]$ model reviewed. Section 7 then looks at the $E^*_2[\delta]$ model, gives the precise decision sets and an approximation of one of these sets corresponding to the canonical cusp catastrophe. Two simple models are then developed which help in illuminating the dynamics of prison riots (Zeeman, Hall, Harrison, Marriage and Shapland, 1976) and, using the model of aggression based on Konrad Lorenz (1963) (Zeeman, 1977), *the cornered rat phenomenon. Section 8 then looks at the* $E^*_3[\delta]$ *model* and, for the set of stationary points on the decision space, derives a local approximation to the canonical Butterfly catastrophe. Here a decision maker mediates between two opposing parties or alternatively we have the above model of aggression when escape is available. Business applications are then briefly discussed with respect to economic modelling and to continuing work in agriculture which started in 1969. These concepts are currently blended with structural models, backed by skeleton 'prime effect' computer models, and used in decision making (Harrison and Quinn, 1977).

2. THE EXPECTED LOSS FUNCTION FORMULATION

2.1. *Beliefs*

Consider a decision maker operating with a belief $F(\Phi\,|\,\delta,u)$, where $\Phi \in \Lambda$ denotes a future outcome, $\delta \in D$ a decision and $u \in \cup$ the environmental variables which may change with time and location. Throughout this paper, unless otherwise specified it will be assumed that the belief distribution function $F \in \mathbf{F}$, the class of distribution functions parameterised by $u \in \cup$, is twice differentiable in Φ and such that the corresponding density $f(\Phi\,|\,u)$ is positive.

2.2. *Utility*

The decision maker is assumed to have a *bounded* Utility function, represented negatively in the way of a Loss function, $L(\delta,\Phi,u)$. That utility functions vary over time is perhaps not universally accepted but is vividly illustrated by Hebron Adams in his reference to the way in which the relative utility of a Kingdom and a Horse change in Shakespeare's 'Richard III'.

2.3. *The Expected Loss Function and Actions*

The routine actions of a decision maker are postulated to be determined by a rule applied to the Expected Loss $E\ [\delta,u]$ where

$$E[\delta,u] = \int_{\Lambda} L\ (\delta,\Phi,u)\ dF\ (\Phi\,|\,\delta,u).$$

2.4. *Definitions*

2.4.1. A Bayes decision δ^* is the infimum of an Expected Loss Function. The set $B(\delta)$ is defined as the set of all Bayes decisions relating to $E[\delta,u]$ over $D \times U$.

2.4.2. The set $S(\delta)$ is the set of all the stationary values of $E[\delta,u]$ and its graph over $D \times U$ is defined as $S(\delta,u)$. In this paper $S(\delta,u)$ will often be the behavioural manifold of an Elementary Catastrophe.

2.4.3. The set of points $C(u)$ will denote the set of points u, for which the corresponding loss $E[\delta,u]$ has two or more local minima.

2.4.4. The set of points $M(u)$ is the set of critical parameter values u for which there are two or more Bayes decisions corresponding to $E[\delta,u]$.

Clearly $M(u) \in C(u)$ and $B(\delta) \in S(\delta)$.

3. UNIMODAL BELIEFS, SYMMETRIC LOSS FUNCTIONS WITH CORRESPONDING MULTI-MODAL EXPECTED LOSS FUNCTIONS

3.1. *The general situation with an example*

In summarizing some of the interesting topics appearing in Smith's thesis and Smith, Harrison and Zeeman (1980), consider the 'Pure Forecasting' situation in which the action of the decision maker does not influence the outcome Φ, (e.g. weather forecasting, small commodity trader etc.), so that $F\ (\Phi\,|\,\delta,u) = F\ (\Phi\,|\,u)$. Suppose also that the Loss Function is bounded and simply a monotonic increasing function of $|\delta\text{-}\Phi|$. If $F \in \mathbb{F}$ is unimodal we can ask if the resulting Expected Loss Function could be multi-modal. The answer is yes, even with familiar standard distributions.

For example, consider $F(\Phi\,|\,u)$ as a Log Normal over $(0,\infty)$ with unit median but variable coefficient of variation k, and a simple double step Loss function

$$L(\delta,\Phi,u) = \begin{cases} 0 \text{ for } |\Phi\text{-}\delta| < b \\ \alpha \text{ for } b \leq |\Phi\text{-}\delta| < c \\ 1 \text{ for } c \leq |\Phi\text{-}\delta| \end{cases}$$

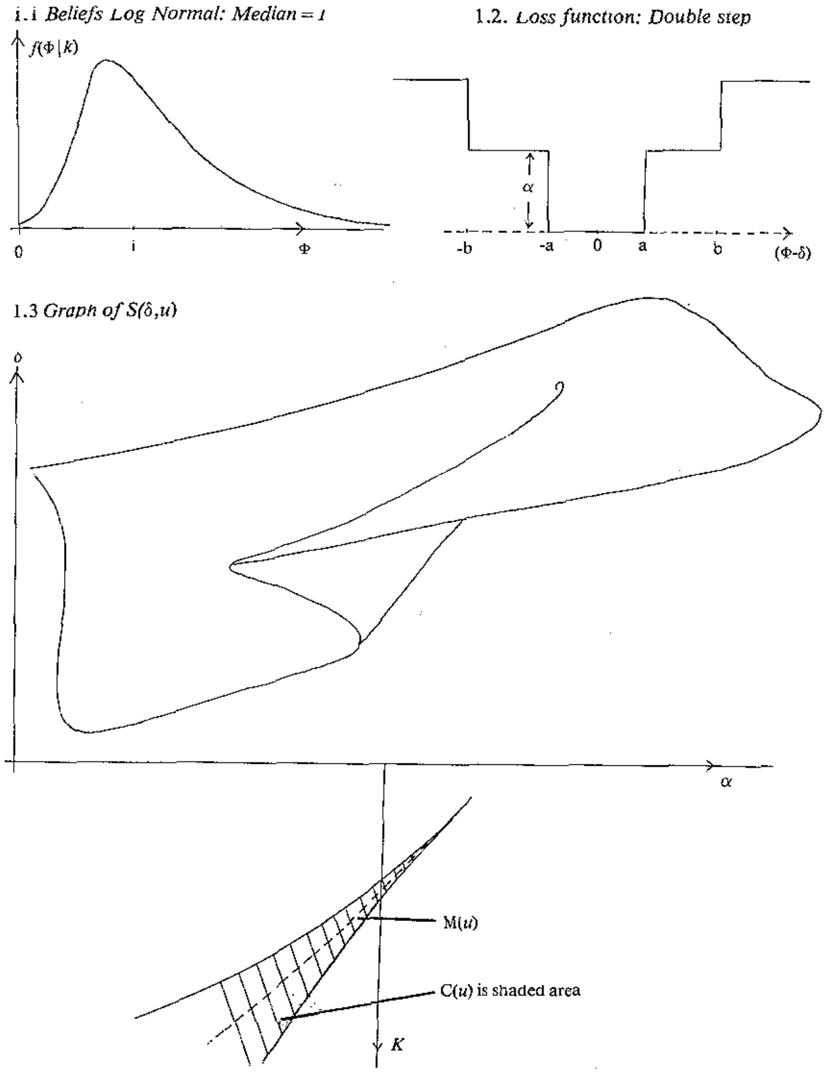
where b and c are fixed but α varies so that

$$u = (\alpha,k) \text{ and } U = (0,1) \times \mathbb{R}^+.$$

Then it can be shown that $S(\delta,u)$ is a cusp catastrophe (Figure 3.1).

FIGURE 3.1

LOG NORMAL EXAMPLE



1.1 *Beliefs Log Normal: Median = 1*

1.2. *Loss function: Double step*

1.3 *Graph of $S(\delta,u)$*

3.2. *An Extension with an example*

The result can be extended outside $\mathbb{F}$ as for example in the following unpublished theorem.

Theorem 3.2.

Let $F(\Phi)$ be a distribution function defined on $(0,\infty)$ such that it is twice differentiable with bounded derivatives. Then, using the previously defined double step loss function with parameter $\alpha \epsilon [0,1]$, if the p.d.f. $f(\Phi)$ is such that

$$f(d\text{-}b) > f(d+b) \text{ for all } d>b$$
$$f(d\text{-}c) > f(d+c) \text{ for all } d>c$$

and if Fisher's Score $\tau(\Phi) = f'(\Phi)/f(\Phi)$ is monotonic over $0<\Phi<2c$, then under the parameterisation $u = \alpha$

(i) the set $S(\delta) = \{\delta : \delta \epsilon [b,c]\}$,

(ii) The Bayes set $B(\delta) = (\{b\};\{c\})$ if $\tau(\Phi)$ is monotonic increasing over (0,2c).

$B(\delta) = S(\delta)$ if $\tau(\Phi)$ is monotonic decreasing over (0,2c)

Corollary: If $f(\Phi)$ is strictly monotonic decreasing with Φ, (e.g. the Pareto Distribution),

(i) $B(\delta) = (\{b\};\{c\})$

(ii) For any parameterisation u, the Bayes decision is

$$\begin{aligned}\delta^* &= c \text{ if } \alpha<\alpha^* \\ &= b \text{ or } c \text{ if } \alpha=\alpha^* \\ &= b \quad \textit{if } \alpha>\alpha^*\end{aligned}$$

where
$\alpha^* = F(2c) - F(b+c)/[F(2c) + F(2b) + F(c\text{-}b) - 2F(b+c)]$

(iii) If $u=\alpha$ then

$$C(u) = \left\{\alpha;\alpha \epsilon \left[\frac{f(b+c)}{f(b+c) + f(0_+) - f(2b)}, \frac{f(2c)}{f(2c) + f(c\text{-}b) - f(b+c)}\right]^r\right\}$$

FIGURE 3.2
THE CRITICAL VALUE OF α^*

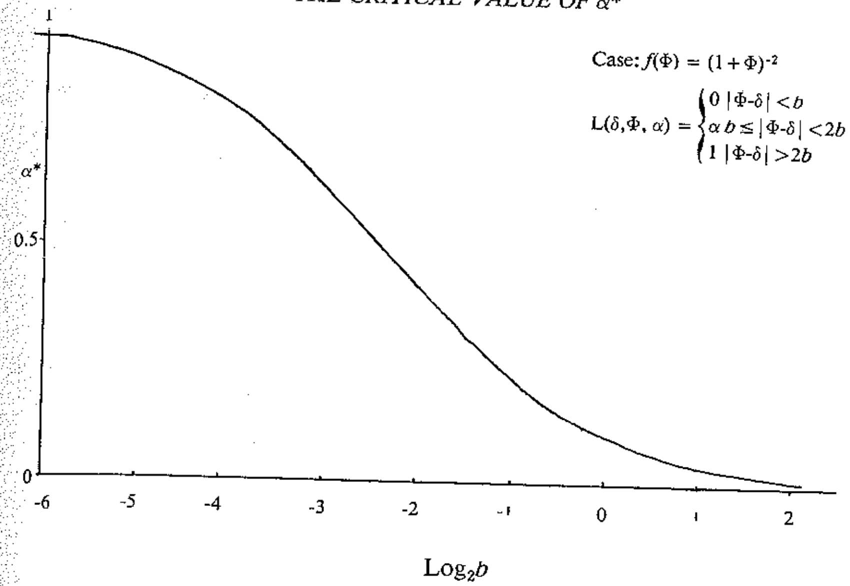
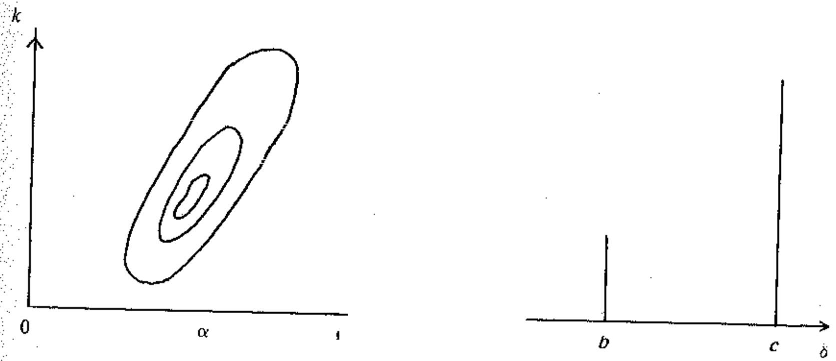


FIGURE 3.3
POLARIZATIONINDECISIONFROMAPOPULATIONWITHNOPOLARIZATIONINBELIEFSORLOSS



Unimodal Population Distribution continuous over $U=(0,1)\times|R^+$ shown by contours of equal p.d.f.

But

Discrete, two point, decision set, polarizing opinion

A particularly interesting case occurs for the Exponential Distribution $F(\Phi) = 1 - \exp(-\beta\, \Phi)$ since Fischer's Score is constant over $(0,\infty)$. What this means is that defining α^* as in the corollary there is just one Bayes decision for each $\alpha \neq \alpha^*$ but corresponding to $\alpha = \alpha^*$ the Bayes decisions comprise the whole interval $[b,c]$. Thus the Exponential Distribution is a critical distribution, with this respect to the Expected Loss function $E[\delta]$, in the sense that it represents a transition in topological types.

Returning to the Corollary, Figure 3.2 shows how, for the Pareto with $f(\theta) = (1 + \Phi)^{-2}$ and $c = 2b$, the value α^* varies with b. There are a number of interpretations of this theorem, one being that if a population of decision makers has a distribution $G(u)$ over $u = (\alpha,k)$ where an individual has a Pareto distribution of beliefs $F(\Phi\,|\,\mu) = 1 - h^k/(\Phi + h)^k$, then even if $G \in \mathbb{F}$ and is unimodal, the population splits into two opposing groups of decision makers with one group G_1 adopting the decision $\delta^* = b$ and the other $\delta^* = c$. This sort of behaviour gives stimulus for understanding how, even with an apparent grouping of beliefs and values, a population can split in its actions. (Figure 3.3).

4. MULTIMODAL FUNCTIONS

We now turn to look at problems involving multimodal information, multimodal loss functions and multimodal expected loss functions.

4.1. *Multimodal Information*

(i) The multi-process models of Harrison and Stevens (1971, 1976) naturally involve multimodal distributions which in the particular case of Dynamic Linear Normal Models (D.L.N.M.) are mixtures of Normal Distributions. These can arise because one is uncertain about an appropriate single D.L.N.M. either over all time or for a particular time interval. Some managers get rather upset over this 'conflict' in information and ask 'how does one reach a decision?' Apart from the uncharitable retort that 'that is what they are paid for', an answer is to specify the loss function and take the corresponding Bayes decision.

(ii) In general hypothesis testing or sampling acceptance a buyer may receive a batch of product from a producer accompanied by the producer 'posterior belief' about product quality. He then carries out his own analyses and perhaps combines or compares the two. It is generally a mistake, and perhaps naive, to adopt the producers posterior as ones own prior. For example if the producers posterior is Normal and ones own tests give a Normal Distribution the result is a single Normal overall posterior showing no conflict. However as a point of interest, if the suppliers posterior is a t-distribution with kernel

$\left[(n-1/n)S_1^2 + \phi^2\right]^{-n/2}$ and ones own distribution is also t with kernel

$$\left[(n-1/n)S_2^2 + (\phi - \mu)^2\right]^{-n/2}$$

then with a simple step loss function $L(\delta,\phi) = 0$ for $|\phi-\delta| \leq A$ and unity elsewhere, the graph of the stationary set of decisions $S(\delta,u)$ represents a Cusp Catastrophe which can be put in canonical form as

Cusp Catastrophe set $S(\delta,u)$
$S(\delta,u) = \{\delta = \psi + \mu/2;\ \psi^3 - b\psi - a = 0\}$ where $u = (a,b)$ and $a = (S_1^2 - S_2^2)\,\mu/4$ $b = (\mu/2)^2 - (S_1^2 + S^2)/2 - A^2$
$M(u) = \{u = (0,b);\ b \geq 0\}$
$C(u) = \{(a,b);\ b > 0,\ 4b^3 \geq 27a^2\}$

Since S_1 and S_2 may be interpreted as the standard errors of individual analyses, the resulting conflict set $C(u)$, which for $A = 0$ gives the straight overall posterior stationary values, is not at all convincing since it is independent of n, the number of observations in each sample! It is also evident that despite the t-distribution converging to a Normal as $n \to \infty$, this way of combining t's does not necessarily converge to the combination of their limiting distributions.

Hence perhaps a better approach is to give the suppliers posterior F_1 a weighting equivalent to n_1 of ones own observations and combine with ones own posterior F_2 based on the equivalent of n_2 observations according to

$$F = w_1 F_1 + (1-w_1)F_2$$
$$\text{where } w_1 = n_1/(n_1+n_2).$$

In this case even if $n_1 = n_2 = n$, $\lim_{n\to\infty} C(u) = M(u)$ which is a set with zero measure. An example of this approach in relation to hypothesis testing is given in Example 5.1 of our 1980 paper.

4.2. *Multi-modal Loss*

(i) A forecaster often meets a multi-modal loss. For example he might have a unimodal belief $F(\Phi | \delta)$ about a future outcome which in combination with his view of the loss function would lead to a decision δ_1*. However, he must report to a Company board who perhaps desire to make a decision in the vicinity of $\delta = \mu$, possibly well away from δ_1* The pressure that such a conflict can exert on the forecaster is not to be underestimated and in reporting his work (which is his decision in his own decision space) he is faced say with a loss function $L = L_1 + L_2$. L_1 reflects his valuation on accuracy, or being true to himself and his profession and L_2 reflects the pressure exerted by what is expected of him in the way of not 'rocking the boat', conforming to expediency, or dogma etc. For example, one of us has been 'unofficially' told prior to a study, that he should not reach a conclusion that would show a plant should not be built and on other occasions both he and colleagues have suffered 'crowding' to the extent that until months or years later when the 'conflicting' view is proved, there was a major penalty to be paid in terms of isolation, abuse and financial loss.

(ii) The above carries across into human affairs particularly in management and industrial relations. For example there are many situations in which a team of people or an individual may desire to operate in one way, say with associated loss L_1, and yet are frustrated by the organisation or a particular manager. To express this frustration by action may involve a penalty with say an associated loss L_2 so there is often no mention of the cause. Clearly as the frustration mounts and toleration decreases a decision is reached to escape physically or mentally, to withdraw co-operation and willingness other than the mimimum, or to confront. An extreme case is the cornered rat phenomena (see 8.2) where in the absence of escape even the weakest will ultimately attack.

(iii) Multi-modal loss often occurs when plans or policies are changed. For example as a short-term forecast for stock control and production planning changes, so this questions whether the production schedule should be altered.

In many instances this has been dealt with by Production Smoothing factors but this is often inappropriate in reflecting the conflict between an aversion to changing the plan reflected by loss component L_1 describing the 'cost effects of such changes', and a desire to choose that schedule which balances stock-out and stock carrying costs reflected by a component L_2. With the production smoothing approach the response of plan to forecast is continuous whereas a more appropriate response often involves an inertial delay in the change of plan until the change in forecast reaches some threshold level.

4.3. *Multi-modal Expected Loss*

(i) Clearly conflicting modes in the Beliefs and/or the Loss function can carry over to the Expected Loss. Furthermore we have shown that even where there are not conflicting modes in either, the resulting Expected Loss can be multimodal.

(ii) Interpersonal behaviour involving the formation and interaction of groups of individuals may often by modelled by multimodal Expected Losses. For example it may happen that a set of individuals each having their own beliefs and losses merge into a group. This may be represented as having a group belief say $(\Phi | \delta) \sim N[\phi_1; V]$ where ϕ_1 represents the 'groups expectation under decision δ and V represents the variation in beliefs. A small value of V may represent high 'internal' group cohesion, certainty, dogmatism, confidence, feeling, etc. whereas a large value may reflect the opposite. Similarly a Group loss may be postulated as

$$L(\Phi,\delta,u) = h[1 - \exp(-(\Phi-\mu)^2/2k)]$$

Here a small k and/or large h may reflect group intolerance to its desires, strong aversions, selfishness, arrogance etc., whereas large k and/or small h may reflect the opposite qualities. The combination of beliefs and Loss function leads to a group Expected Loss $E_1[\delta]$ with one minimum. We can now consider the interaction between interrelated and perhaps opposing groups. For two groups there are various approaches.

Each group may be looked at separately with the interconnections being expressed as an additional loss function component.

A sudden change in action on the part of one group seems to change the parameter values in the others Expected Loss, thus mimicking the role of the trigger mechanism that makes the heart beat (Zeeman, 1977). Alternatively the groups may be combined with the decision vector being mapped onto $\mathbb{R}^2$ or the two Expected losses may be merged into a combined expected loss, for example,

$$E(\delta) = \alpha_1 E_1[\delta] + \alpha_2 E_2[\delta]$$

and the two groups considered in a combined fashion.

4.4. *Duality*

The same $E[\delta]$ and hence its stationary points and Bayes decisions can arise in many ways. We show two particular cases as duals:

(i) Take a multimodal belief with p.d.f.

$$\Sigma_{i=1}^{n} w_i f_i(\Phi)$$

and a loss function $L(\delta;\Phi)$
so that

$$E[\delta] = \Sigma_{i=1}^{n} w_i \int L(\delta;\Phi)\, f_i(\Phi)\, d\phi$$

Suppose that a bijective mapping $T_i: \Phi \rightarrow \theta$, $i = 1 \ldots n$
exists such that $f_i(T_i^{-1}\theta)\,|d\theta/d\Phi| = f(\theta)$
then $E[\delta] = \Sigma_{i=1}^{n} w_i \int L(\delta; T_i^{-1}\theta) f(\theta)\, d\theta$
which gives the dual problem as

(ii) An outcome θ with a p.d.f. $f(\theta)$ and a multimodal loss function

$$\Sigma_{i=1}^{n} \mathrm{w}_i L(\delta;\ T_i^{-1}\theta)$$

leading to the same $E[\delta]$

5. SMITH'S THEOREM

Smith (1978) in his thesis, and in his paper on the Mixture of Distributions (1979) gives two general theorems relating to particular potentials. One of these is also given in our 1980 paper and the relevant extract is reproduced here for convenience.

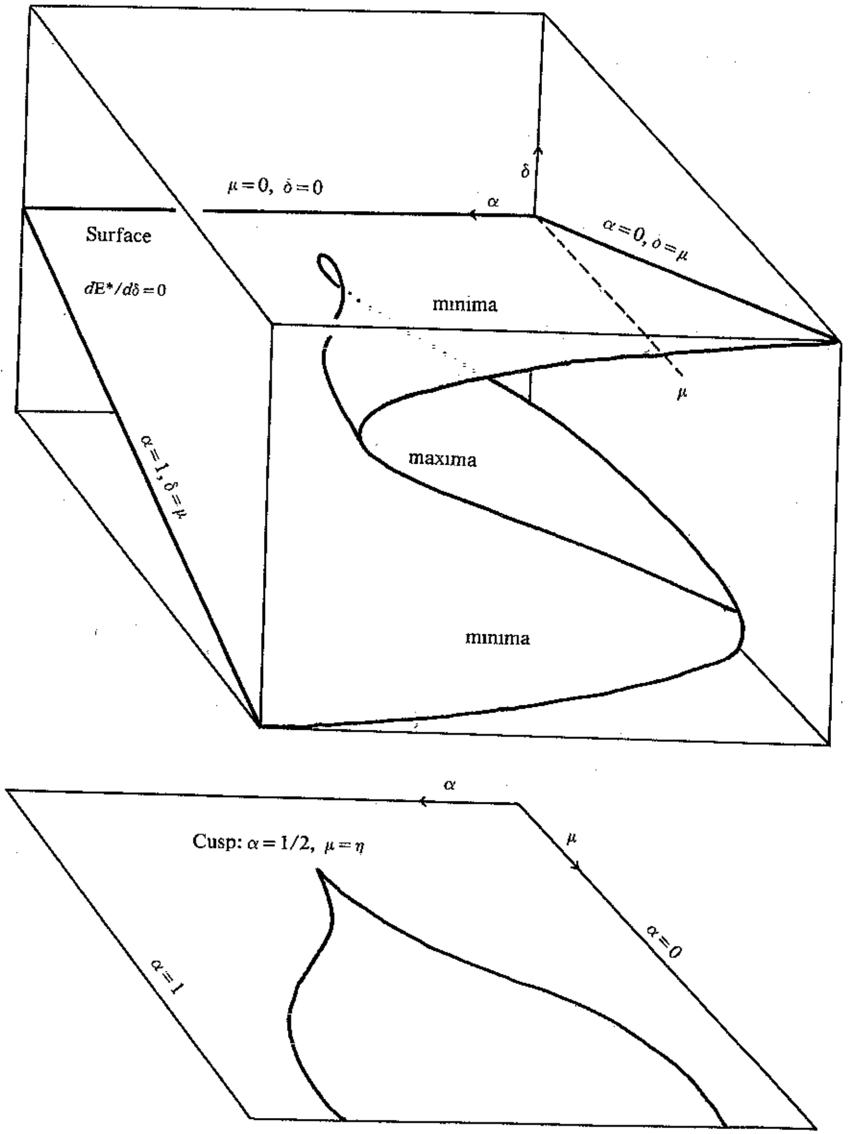
5.1. *A Cusp Catastrophe in Bayesian Decision Theory*

A key use of the geometry of the cusp catastrophe in Bayesian estimation theory is given in the following theorem. Given a function E of s, let E', E'', E''' denote respectively the first, second and third derivatives of E with respect to s. We say $E(s)$ is of type T if it is C^∞, symmetric, strictly increasing in $|s|$, with $\lim_{s \rightarrow \infty} E(s) = 1$ and satisfying the three conditions

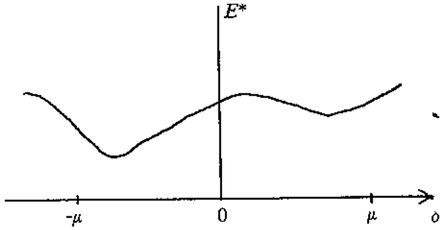
i) E'' has one zero in $(0,\infty)$ at η, say
ii) E''' has one zero in $(0,\infty)$ at λ, say
iii) the images of $(0,\eta)$, (η,λ) under the function E''/E' have empty intersections.

For example $E(s) = 1 - exp(-\frac{1}{2}s^2)$ is the type T with $\eta = 1$ and $\lambda = \sqrt{3}$.

FIGURE 5.1
E EXHIBITS A CUSP CATASTROPHE OVER* (α,μ)



Theorem 5.1



Let E be of type T. Let

$$E^*(\delta) = \alpha E(\delta + \mu) + (1\text{-}\alpha)E(\delta\text{-}\mu)$$

defined over the two dimensional parameter space given by $0 < \alpha < 1$ and $\mu > 0$. Then E*(δ) exhibits one unique cusp catastrophe whose coordinates are given by

$$(\delta,\alpha,\mu) = (0,\ \tfrac{1}{2},\eta)$$

with normal factor α and splitting factor μ.

The resulting surface of stationary values of E* is illustrated in Figure 5.1

6. MIXTURES OF NORMALS AND CONJUGATE NORMAL LOSS FUNCTIONS

6.1. *Introduction*

6.1.1. Beliefs

In the remainder of this paper the belief distributions are either Normal or mixtures of Normals so that

$$F(\Phi\,|\,\delta,\text{u}) = \Sigma_{i=1}^{n} w_i N_i (\Phi\,|\,\delta,\text{u})$$

where N_i $(\Phi\,|\,\delta,\text{u})$ represents a Normal Distribution where the mean θ_i, the variance V_i, and the relative weight w_i may all be functions of u and δ. Naturally $w_i \geq 0$ and $\Sigma_{i=1}^{n}$ $\text{w}_i = 1$. For simplicity, we will consider only univariate Normals so that $\Phi \in \mathbb{R}$.

6.1.2. *Loss Function*

Following Lindley (1976), for the reason that theoretical insights into problems are facilitated by adopting conjugate Loss functions, the general loss function considered is now of the form

$$L(\Phi,\text{d},\text{u}) = \Sigma_{i=1}^{m} L_i(\Phi,\delta,\text{u})$$

where

$$L_i = h_i\,(1 - \exp(-(\Phi\text{-}\psi_i)^2/2k_i))$$

and $0 < h_i < \infty$, ψ_i and k_i may be functions of δ and u.

6.1.3. *The Expected Loss Function*

The resultant Expected Loss function is then

$$E[\delta,\text{u}] = \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} \alpha_{ij} E_{ij}\ [\delta,\text{u}]$$

where

$$E_{ij} = 1\text{-}(k_j/(k_j + V_i))^{1/2} \exp\ (-(\theta_i\text{-}\psi_j)^2/2(k_j + V_i))$$

6.2. *The n- Conjugate Normal Expected Loss Function E_n*[δ]*

Let $E_i[\delta,\text{u}] = 1 - (k_i/(k_i + V_i))^{1/2} \exp -((\delta\text{-}\mu_i)^2/2(k_i + V_i))$

where $(k_i, V_i,\mu_i) \in \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}$,

$\alpha_i > 0$, and define

$$E_n{}^*[\delta,\text{u}] = \Sigma_{i=1}^{n} \alpha_i E_i[\delta,\text{u}]$$

A $E_n{}^*[\delta]$ model is then one which involves the above type of Expected Loss Function

6.3. *An E_1*[δ] Model with Conflicting Decisions*

It is interesting to refer first to the case in which $n = 1$.

Consider

$$(\Phi\,|\,\delta, u) \sim N\ [c + \delta; V\ ]$$

$$L\ (\Phi,\delta,u) = h\Big[1 - \exp\ [-(\Phi\text{-}\mu)^2/2k]\Big]$$

$$\text{so that } E[\delta,u] = h\ [\ 1 - \left(\frac{k}{k + V}\right)^{1/2} \exp\ (\frac{-(\delta\text{-}(\mu\text{-}c))^2}{2(k + V)}\ )]$$

It is very clear that if $h,k,V,$ and c are all constant then there is one and only one Bayes decision at $\delta^* = \mu\text{-}c$. However, in practice decision makers do not generally adopt a decision which varies continuously with $(\mu\text{-}c)$ and it was desired to examine whether the sudden changes in decision could be captured naturally by making more realistic assumptions. Often V and k are not constant, in that the uncertainty associated with an outcome increases as the decision departs from the familiar or status quo. Similarly k, the tolerance to errors, often decreases as the decision involves increasing costs. Our 1980 paper shows, using an investment example, that such dependencies of V and/or k on δ can lead to conflicting decisions, to delayed decisions and to discontinuous changes in the Bayes decision, as (μ,c) changes continuously. Hence the assumption of the constancy of V and k is very unstable in that the slightest perturbation in this assumption can lead to totally different qualitative behaviour on the part of a decision maker who apart from acting conservatively and continuously now also makes discontinuous and qualitatively different decisions in response to small continuous changes in the environmental variables.

7. THE $E_2{}^*$ [δ] MODEL

7.1. *An E_2 [δ] Model with associated Decision Sets*

Using a trivial extension of Smith's Theorem stated in Section 5, consider the case in which for $E_2{}^*[\delta,u]$ $k_i = k$ and $V_i = V$ $(i=1,2)$ are constant, without loss in generality $\mu_1 = -\mu_3 = \mu$, and the environmental variable is $\cup = (r,\mu)$ over $(0,1)\mathrm{x}\ \mathbf{R}$, where $r = \alpha_1/(\alpha_1+\alpha_2)$. Then it is found that $E_2{}^*$ [δ] exhibits cusps only along the co-ordinates

$$(\delta,r,\mu) = (0\ ;\ \tfrac{1}{2}\ ;\ (k+V)^{1/2})$$

with, in Catastrophe terminology, normal factor r and splitting factor μ. There are two cusp points over $\cup$ according to the sign of $(k+V)^{1/2}$ and, because of the symmetry, the graph $S\ (\delta,u)$ exhibits what may be regarded as two 'back to back' cusp catastrophes (Figure 7.1). Writing $\phi(x)$ as the Standard Normal density, $S\ (\delta) = \{\delta = (k+V)^{1/2}d; \alpha_1(d-\mu)\phi(d-\mu) + \alpha_2(d+\mu)\phi(d+\mu) = 0, u \in \cup\}$

$$C(u) = \{u = (r,\mu)\ ;\ 1\text{-}g \leq r \leq g,\ g = 1 + c\ \exp\ ((1/c\text{-}c)/2), c = \mu + (\mu^2\text{-}1)^{1/2}\}$$

$$M\ (u) = \{u = (\tfrac{1}{2},\mu)\ ;\ \mu^2 > k+V\}$$

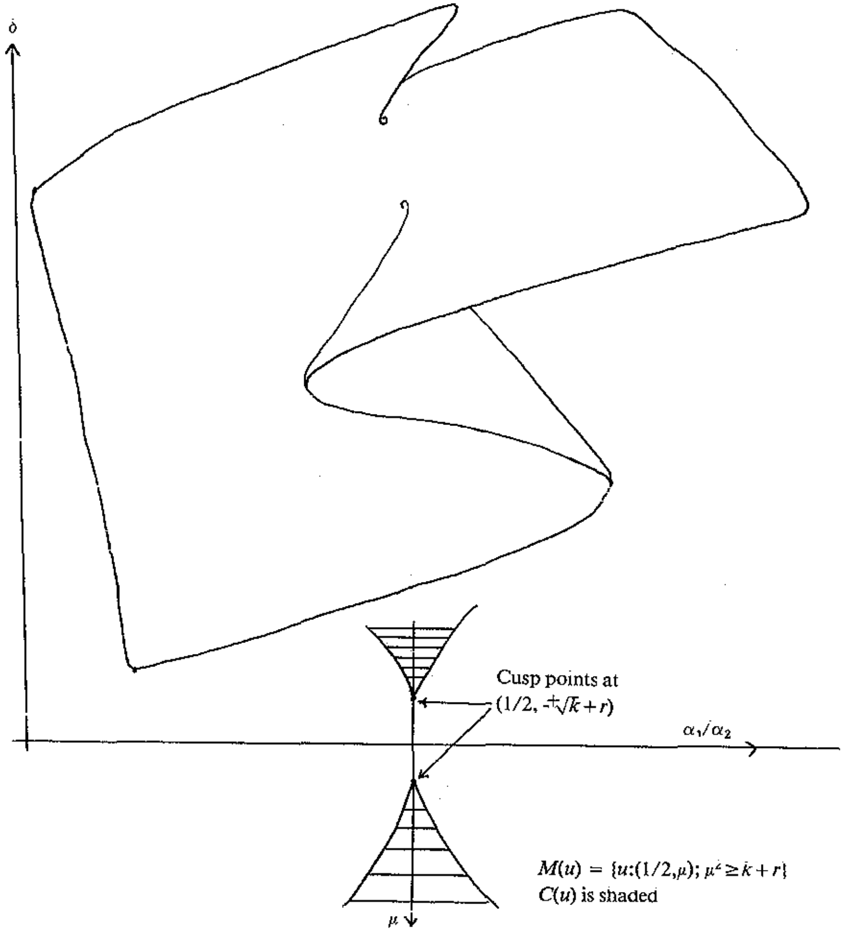
Measuring μ and δ in units of $(k+V)^{1/2}$, or without loss, letting $k+V=1$, then a mapping T: $\cup \rightarrow W$, which gives quite a good approximation to the canonical cusp catastrophe is as follows:

Approximate Mapping to the Canonical Cusp
$S\ (\delta,w) = \{(\delta,a,b); \delta^3\text{-}b\delta\text{-}a = 0,\ (a,b) \in W\}$ $b = 3[\mu^2\text{-}1]$ as splitting factor, $a = 2\log\ (\alpha_1/\alpha_2)$ as normal factor

Thus conflict can only occur if $\mu^2 > 1$, and for fixed a, the larger μ^2, the greater apart are the conflicting decisions.

FIGURE 7.1

$S(\delta,u)$ FOR THE E_2^* [δ] MODEL



7.2. *A simple E_2^* model for Institutional Disturbances*

In our paper on institutional disturbances (Zeeman, Hall, Harrison, Marriage and Shapland, 1976) we were specifically concerned with ways of viewing prison riots. The resulting model used hypotheses that

(i) the graph of the 'states' in the prison could fruitfully be looked upon as a Cusp Catastrophe with Tension as a Normal factor, Alienation as a splitting factor and the amount of disturbance as the behavioural factor;

(ii) there was a natural dynamical flow on the cusp manifold leading away from quiet states and also from 'abnormal' disturbance states;

(iii) locally a stationary stochastic process described the departures away from the manifold.

Let us now consider a simple model based upon the foregoing aiming to capture the main aspects. Being a coercive environment with virtually no escape in the short-term we may consider the prisoners group beliefs about the outcome Φ of taking action as

$$(\Phi\,|\,\delta) \backsim N\,[c+\delta;V\,]$$

Their Loss function is bimodal reflecting, on the one hand, their desire to change their circumstances and overtly show their discontent and, on the other hand, their aversion to penalties if their behaviour departs from the imposed norm. Take the resulting Expected Loss as

$$E_2^*[\delta,u] \;=\; \Sigma_{i=1}^{2}\,\alpha_i\,[1 - (k_i/(k_i+V_i))^{1/2}\exp\left(-(\delta-\mu_i)^2/2(k_i+V_i)\right)].$$

Now consider the simple case in which $k_1=k_2$, $V_1=V_2$ and without further loss in generality $k_1+V_1=1$ and $\mu_1=-\mu_2=\mu\geq 0$. Clearly the first two equalities would not hold generally nor be constant since tolerances, grouping and variation in beliefs will change during any escalation of disorder. However, to a first approximation, and for a simple qualitative insight, this variation can be tranferred to variation in α_1 and α_2. Hence α_1/α_2 is here interpreted as the relative strengths or tension of the conflicting desire and aversion to overt demonstration and μ as the alienation representing the compatibility or the 'emotional distance' between the restricted activity of prisoners as imposed by the authorities and their currently desired activities. It then follows from the previous results on the E_2^* $[\delta,u]$ model that $S\,(\delta,u)$ is a Cusp Catastrophe manifold with Normal Factor proportional to Tension and Splitting Factor Proportional to Alienation.

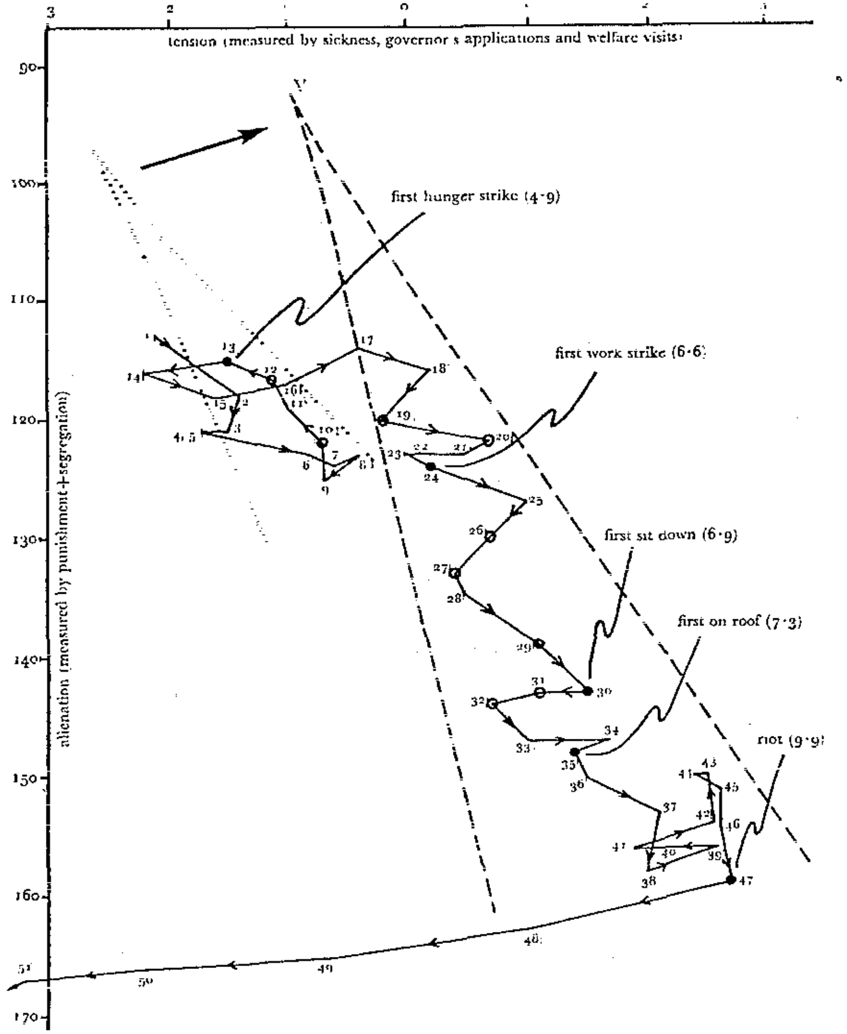
The practical problem of obtaining measurements which manifest tension and alienation is discussed in the 1976 paper. The tension measure was based on such factors as sickness, and as such is a measure of total rather than relative tension. Hence using this measure, after the first group disorder in week 13, the accompanying tightening of authorative control with perhaps increased penalties can be interpreted to mean that for the relative tension to reach the same threshold level precipitating overt group action the total tension, as reflected in the measurement, would have to increase. This is an interpretation of why on the observed environmental space the scene of the action changes markedly after the first group disorder before settling into the expected zig-zag escalation of conflict which lead to the destruction of half of prison (Figure 7.2). The E_2^* model has since lead to an additional way of monitoring the state of the prison in which α_1/α_2 represents the relative power of prisoners and it is hoped that this can be published later.

The more general case in which k_i and V_i differ and vary has been examined in unpublished work. The effect of a decrease in k_1 is similar to that of a decrease in V_1 and qualitatively the same as an increase in h_1 and hence α_1, although quantitatively different. For example, whereas in our simple model, allowing α_1/α_2 to vary whilst keeping their sum constant results in $C(u)$ being symmetric about a fixed ratio (here ½), the variation in the ratios of the k's and V's disturbs that symmetry.

For an approach to the dynamic modelling of flows on Catastrophe manifolds using multi-process models reference may be made to Sawitzki (1978).

FIGURE 7.2

A MODEL FOR PRISON DISTURBANCES



Analysis of Gartree data for 1972. Time path of tension and alienation is plotted weekly throughout the year (numbers indicate weeks). The serious incidents are indicated by circles. The solid circles indicate those incidents involving nearly all the inmates in a new form of mass protest; the numbers in brackets indicate an assessment of seriousness (out of 10). A possible initial position of the cusp is shown dotted an a possible subsequent position is shown dashed; the movement of the cusp may represent a higher tolerance level of tension in the institution after the first mass protest.

7.3. *Aggression and the Cornered Rat*

The foregoing institutional disorder example may be likened to Zeemans (1977) well known catastrophe model of Konrad Lorenz's (1963) statement that rage and fear are conflicting factors.

The prisoner may be compared to a dog or a rat which is cornered, tormented and has no chance of escape. The conflict may again be modelled through the Loss function in which there is the desire to retaliate to the torment by attacking and yet an aversion to the consequences to attack. This results in an E_2* $[\delta,u]$ model and taking $u = (\alpha_1/\alpha_2,\mu)$ as for the prisoner we have $S(\delta,u)$ as the Cusp Catastrophe.

In this 'cornered version' of Zeemans example α_1/α_2 may be taken to measure the ratio of rage to fear, in the sense of the desire to relieve the torment compared to the aversion to the consequences of attack, and μ as measuring the amount of emotional disturbance generated by the incompatibility of desire and aversion. The decision δ refers to the behaviour of the dog and generally falls into one of the two qualitatively different states 'attack' or 'do nothing'. Within these states δ varies according to the way in which the dog accepts its torment or according to the intensity of its attack. If the dog is a Bayesian Bloodhound and the tormenting gradually increases then it switches to the attack when $\alpha_1=\alpha_2$, (i.e. $M(u) = \{u;\alpha_1=\alpha_2\}$.) no matter how badly it is matched with its opponent. The example is perhaps not irrelevant to industrial relations, in those situations in which a work force experiences, a recession with greatly reduced orders, a long serving work force, no alternative employment opportunities, insufficient compensation, mounting rationalisation and meaningless tasks. Here the response may at first be calculative in terms of some acceptance of redundancy terms, then resistance to management change with non-cooperation leading perhaps finally to a sit-in. This powerless response is very different to that of a powerful aggressive confident work force who themselves force the changes. Both aspects can be captured together by the back to back cusps with appropriate translations of the resultant δ* actions dependent upon the environment u.

8. THE E_3* MODEL

8.1 *Approximations to the Decision Sets of an E_3* Model*

Smith (1978), (1979) developed a further theorem, which includes the E_3* Model as one case. Within the E_3* class a particular model is selected which, for those people used to catastrophe modelling involving the Butterfly Catastrophe, can give insight into the behaviour of decision makers.

Again we will take $k_i=k$ and k_i+V_i = constant, in this case for scaling convenience $k+V = 1/3$, $(i=1,2,3)$, and we appeal that variation in these quantities is, to the first order, captured by variation in the α_i. Thus

$$E_3^* [\delta,u] = \Sigma_{i=1}^{3} \alpha_i E_i [\delta]$$

where

$$E_i [\delta,u] = 1-(3k)^{1/2} \exp(-3(\delta-\mu_i)^2/2).$$

Consider the case in which $\mu_1 > \mu_2 = 0 > \mu_3$. Define $\mu > 0$ by $\mu_1-\mu_3 = 2\mu$ and the relative weights r_i by

$$r_i = \alpha_i / \Sigma_{j=1}^{3} \alpha_j \text{ so that } \Sigma r_i = 1$$

Then the application of Smith's Theorem shows that there is a unique Butterfly point at the co-ordinate

$(\delta,r_1,r_3,\mu_1,\mu_3) = (0,r,r,1,-1)$

where $r = [2(1+2\exp(-3/2)] = 0.346$.

The graph of $S(\delta,u)$ may be obtained readily by computer calculation. Here is given a local approximation to it by expanding as a Taylor Series around the Butterfly point and expressing the result as the canonical Butterfly catastrophe so that the nature of the Normal, a, Splitting, b, compromise, c, and Butterfly, d, factors may be seen. The geometry of the canonical butterfly is described in Poston and Woodcock (1974) although the reader should notice the sign reversal of the four factors.

Local approximation to $S(\delta,u)$ around Butterfly Point $(0,r,r,-1,1)$

$$S(\delta,(a,b,c,d)) = \{(\delta,a,b,c,d);\ \psi^5-d\psi^3-c\psi^2-b\psi-a=0\}$$

where

$$\delta = \psi + [3[r_3-r_1] + 2[\mu_3 + \mu_1]]/6$$
$$d = 10\,[(r_2-0.31) + 2(\mu-1)/3]$$
$$c = -10\,[\mu_3 + \mu_1]/3$$
$$b = -7\,[r_2-0.31]$$
$$a = 10\,[2[\mu_3 + \mu_1] - 3[r_3-r_1]/27]$$

As might be expected the bias factor, c, depends only on the imbalance in the relative distances μ_3 and $|\mu_1|$. whereas the normal factor, a, also includes the imbalance in the relative weightings r_3 and r_1. The splitting factor depends only upon the 'middle' relative weighting r_2 whereas the Butterfly factor additionally, depends upon the distance $(\mu_3-\mu_1)$.

8.2. *Simple Models Relating to Decision making in the face of conflicting interests.*

8.2.1. *Hierarchies and Arbiters*

Most hierarchical command chains can give rise to conflicts which might be viewed in the light of an $E_3^*[\delta]$ model. For example consider the case in which a manager is pressurised from above to take decisions according to one criterion or viewpoint, from below according to another and yet he has his own criterion and beliefs which fall somewhere in the middle. Such situations may be charged as the economic environment worsens and the conflicting demands become strongly opposed. An arbiter is often faced with a similar type of problem. Taking the above model in the vicinity of the Butterfly point with say $\mu_3 + \mu_1 = 0$, first look at the evenly balanced case in which $r_3 = r_1$, so that the manager is faced with a symmetric $E_3^*[\delta]$ and so that $a = c = 0$, $b \propto (r_2-0.31)$ and $d \propto (r_2-0.31) + 2(\mu-1)/3$. Then the Bayes decisions δ^* are

(i) $\delta^* = 0$ if $(d<0, b<0)$

or if $(d>0, 16b+3d^2 \leq 0)$

which is the 'middle course' decision

otherwise

(ii) $\delta^* = \pm(d + |(d^2+4b)^{1/2}|)/2$

where there are two conflicting extreme Bayesian decisions

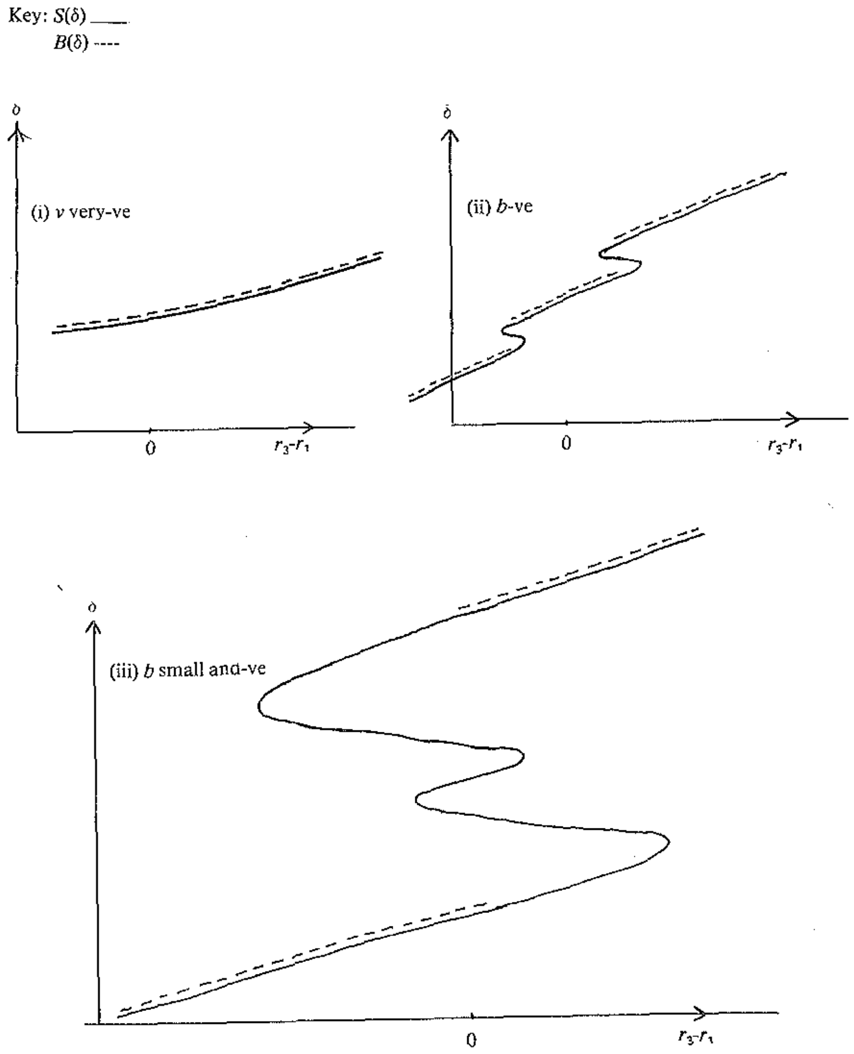
Holding b constant and slightly negative, d constant and positive, but allowing r_3-r_1 to vary around zero, three examples of the stationary and Bayesian decisions are shown in Figure 8.2.

8.2.2. *Aggression*

Considering the example of the dog (7.3), but this time with a possible escape, we might map the behaviour into the decision space $\mathbb{R}$ so that in the vicinity of $\delta = 0$, acceptance of the torment is indicated, δ distinctly negative indicates escape with perhaps varying degrees of rapidity and δ distinctly positive is indicative of attack with varying degrees of intensity. If this is modelled in an E_3* fashion with the third component reflecting the utility of escape then the main features of the dogs behaviour are captured, possibly more effectively than in Zeemans (1977) original cusp model.

FIGURE 8.2

BUTTERFLY CROSS SECTIONS FOR THE $E_3^*[\delta]$ MODEL



8.3 *Systems Applications of E_3* models*

In macro modelling unwise expedient control action which purports to stabilize conditions often does just the reverse. This is rather like trying to control a swinging pendulum by delayed responses which push the pendulum in the direction in which it is naturally travelling. Such action often destroys rather than maintains Equilibrium.

In two of the main macro-modelling situations with which one of us is involved catastrophe cross sections are very useful in providing insight, maps and monitoring devices. As a flavour of the way in which these are used in conjunction with systems models figures 8.3 and 8.4 are shown. The two applications shown here which were first phrased as threshold models in the early 70's provided the main motivation for the study of catastrophe theory and to the development of the decision theoretic formulation described earlier in the paper.

The first figure 8.3 considers a measure of a farmers perception or 'expectation' relating to the decision on what price to pay for a heifer. For simplicity, here the decision is judged on the basis of his realisation one year later taking into account the deflated value of the fat cow and the calf output price. Consequently we have a Butterfly type of cross section with delayed decisions which illustrates perception or expectation against actuality. The contradictions arise from the unanticipated feedback and delayed dynamical responses to central action and farmer action and are sustained by desire, aversion and political propoganda. The breaking of the central equilibrium level is predictable and the almost unstoppable dynamical response magnified by out of phase 'control action' ensures the catastrophic rise in returns followed by the disastrous collapse and consequent ruin of many beef producers (Harrison and Quinn, 1977). This is particularly relevant as we approach the critical 1979 agricultural decisions in the E.E.C. which, in the face of the world agricultural situation, the pressures arising from other agricultural sectors and political lobbies, debt, and the remote reductionist way of dealing with the 'control' of strongly interrelated systems, is very likely to lead to actions which will heighten the consequences and have similar disastrous effects on beef in 1982/3, earlier on some other sectors and later on still others. After all as one official said 'You can never be sure of the future and we will have to see whether it happens again'. Currently this sort of diagram is used as an invaluable monitor and means of communication relating to a systems model of agriculture which covers numerous sectors and associated industries. As a means of communication, it helps to capture some of the main phenomena arising out of the interrelationships. As a monitor, it is taken as a source of confirmation or disagreement with a stated view of the future.

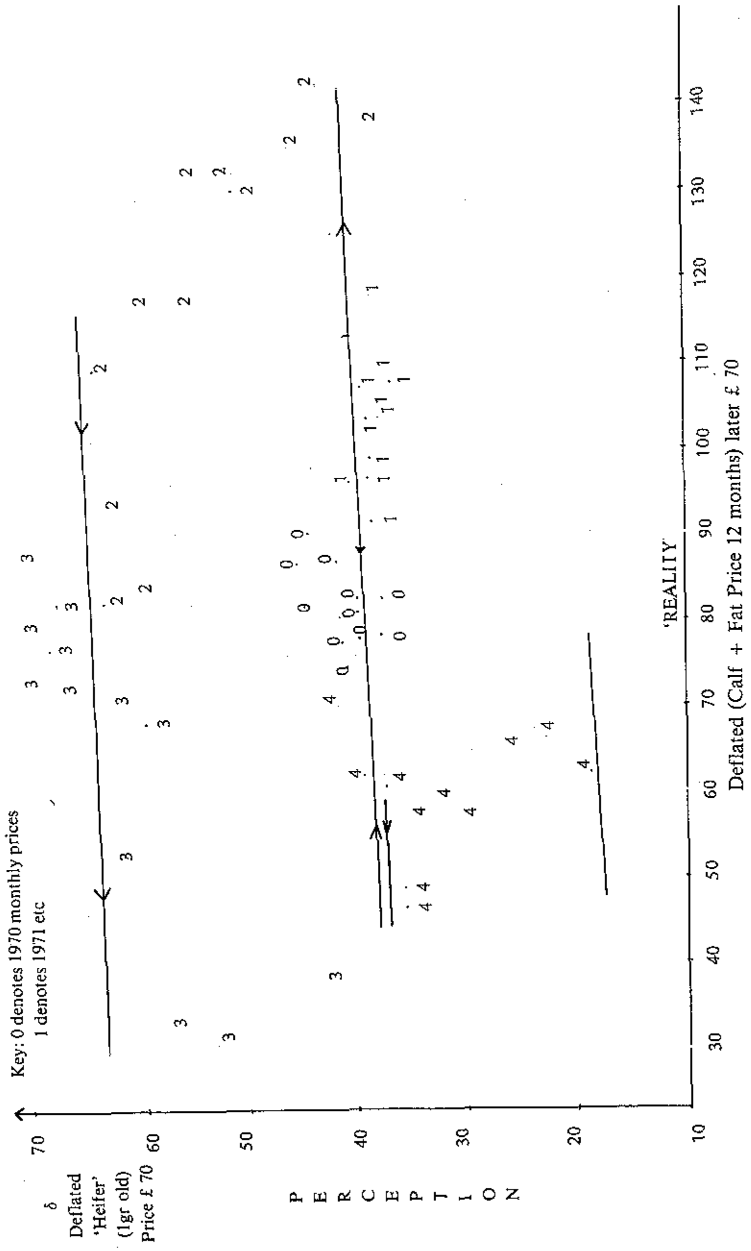


FIGURE 8.3

THE BEEF BREEDER'S DECISION: BUTTERFLY 'TYPE' CROSS SECTION OF · PERCEPTION VERSUS ACTUALITY', OF THE DESTRUCTION OF MARKET BALANCE MONTHLY PRICES 1970-74

Figure 8.4 shows one phase plane diagram constructed in 1971 relating to work on Cash Flow and Economic modelling. This is just one important extract from the overall systems view which lead to a forecast issued in 1971 that in 1974/5 the U.K. would experience the worst depression since the 1930's. The approach relating to this particular figure concerned the government as a decision centre receiving delayed and screened information, with a major utility component relating to retention of power. Two of the main environmental 'conflicting' variables on this centre are illustrated in fig. 8.4 as Unemployment and Debt in the form of Balance of Payments. (This can be thought of as an approximate canonical Butterfly model, with the difference of these conflicting variables being the Normal factor and their sum the Splitting factor). The expedient decisions in exchanging pressure arising from one of these variables for delayed but more intense pressure from the other are clearly seen. Phase I is defined as the period 'U.K. internal conflict' which in 1967 transformed into Phase II 'conflict between developed nations' and then in 1975 to 'General economic conflict'. In Phase I the three main types of decision were reflate, relative inaction and deflate (Fig. 8.5). As unemployment increased with positive balance of payments the expedient decision involved reflation. Since, among other things, reflationary action was taken when industry was geared for the existing deflationary period and since there is significant delay in responding with more finished goods needing more imported raw materials, it was hardly surprising to find that natural dynamical responses turned the problem into one of debt leading to expedient deflation. This in turn hit industry when it was geared for a reflationary period so that supply exceeded home demand. Thus product dumping abroad together with the reduction of imported raw materials and finished product, resulted in a healthy trade balance but a delayed major rise in unemployment. The fact that this cosy internal U.K. drift to ever worsening conditions was rudely affected by external situations in 1967 is very clearly indicated on this phase diagram. The simultaneous pressure from the two sources signalled a qualitative change in environment labelled as Phase II which in encouraging the 'sure-thing losing gamble', the mad dash for growth in 1972, ensured the economic imbalance resulting in Phase III. It is interesting to hear the comfortable explanations of many people that Phase III occured because of the oil situation and that the agricultural crisis arose because of Russian grain buying. They seem to completely miss the point that both these occurrences might have been just single important manifestations of deeper phenomena.

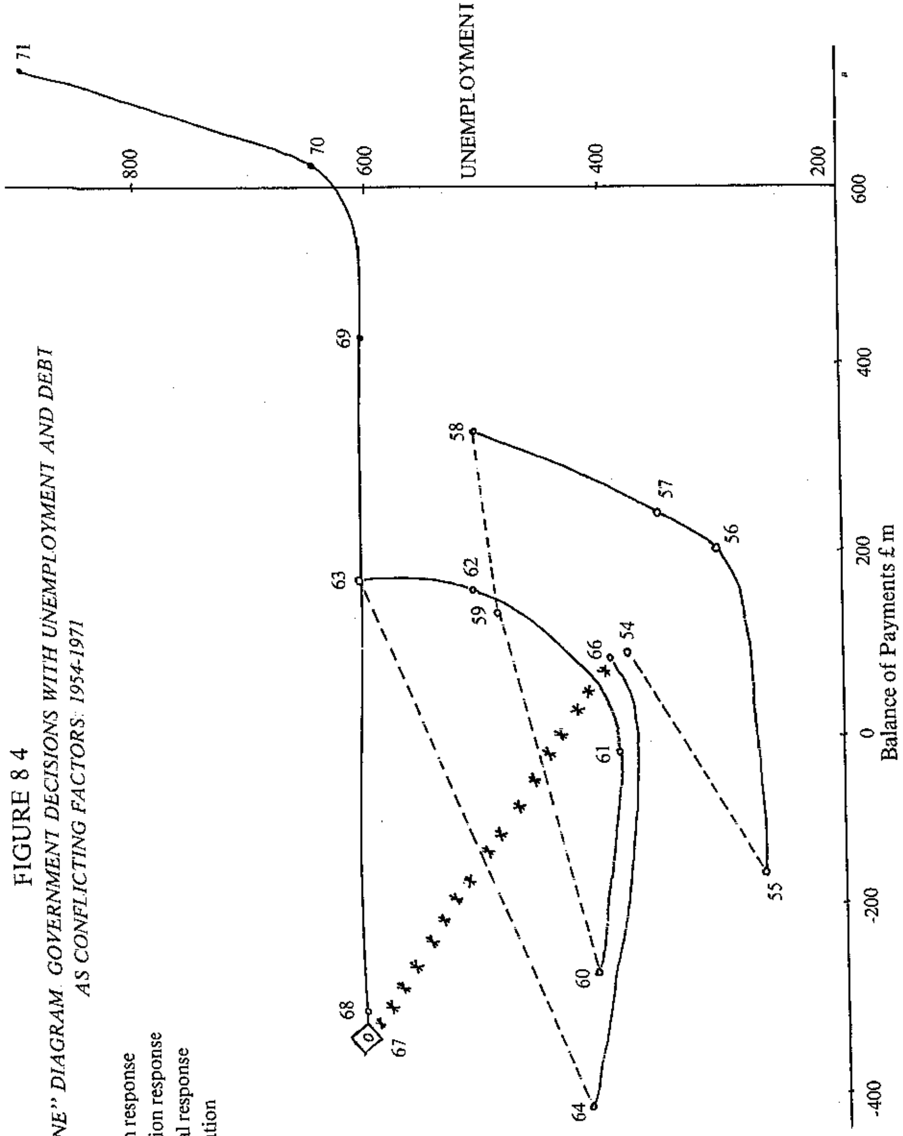
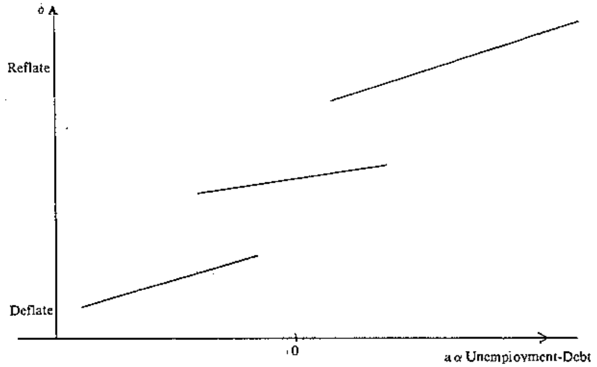


FIGURE 8 4
"PHASE PLANE" DIAGRAM. GOVERNMENT DECISIONS WITH UNEMPLOYMENT AND DEBT AS CONFLICTING FACTORS. 1954-1971

Key --- reflation response
____deflation response
*** external response
⊙ Devaluation

FIGURE 8.5
CROSS SECTION: DECISION δ, AND NORMAL FACTOR



REFERENCES

HARRISON P.J. and STEVENS C.F. (1971). A Bayesian approach to short-term forecasting. *Oper. Res. Quart.* **22**, 341-62.

— (1976). Bayesian forecasting (with discussion). *J. Roy. Statist. Soc. B* **38**. 205-247.

HARRISON P.J. and QUINN M.P. (1977). A brief description of the work done to date concerning a view of agricultural and related systems. In *Econometric models presented to the Beef-Milk Symposium.* E.E.C. Agriculture. EUR 6101 EN.

LINDLEY, D.V. (1976). A class of utility functions. *Ann. Statistics* **4**, 1-10.

LORENZ, K. (1963). *On aggression.* London: Methuen.

POSTON, T. and STEWART, I.N. (1978).*Catastrophe Theory and applications.* London: Pitman.

POSTON, T. and WOODCOCK, A.E.R. (1974). A geometrical study of the elementary catastrophes. *Lecture Notes in Mathematics* **373.** Berlin: Springer-Verlag.

SAWITZKI, G. (1978). *Bayes-statistik für stochastische prozesse mit nicht-linearer struktur.* Doktors dissertation, Ruhr-University, Bochum.

SMITH, J.Q. (1978). *Problems in Bayesian statistics related to discontinuous phenomena, catastrophe theory and forecasting.* PH.D. thesis, Warwick University.

— (1979). Mixture catastrophes and Bayes decision theory. *Proc. Cambridge Philos. Soc.* **86**, 91-101.

SMITH, J.Q., HARRISON, P.J. AND ZEEMAN, E.C. (1980). The analysis of some discontinuous decision processes. *European J. of O.R.*

THOM, R. (1972). *Stabilité structurelle et Morphogenese.* New York: Benjamin. English translation (1975) by D.H. Fowler.

ZEEMAN, E.C., HALL, C.S., HARRISON, P.J., MARRIAGE, G.H. and SHAPLAND, P.H. (1976) A model for institutional disturbances, *Br.J. of Math. and Stat. Psychology* **29**, 66-80.

— (1977). A model for prison disturbances. *Brit. J. of Criminology* **17**, 251-263.

ZEEMAN, E.C. (1977) *Catastrophe Theory, selected papers 1972-1977.* New-York: Addison-Wesley.

DISCUSSION

S.E. FIENBERG (*University of Minnesota*):

The three papers presented at this Session have been linked together under a common title. In fact they are only loosely related to one another. The key links would appear to be (1) between the Makov paper on sequential learning, and Section 6 of Smith's paper on change-point problems, (2) the use in all three papers of the idea of recursive updating (this appears explicitly in the Makov and Smith papers, and only implicitly in the Harrison and Smith paper via the use of smoothed data for the tension and alienation variables in their prison riot example of Section 7). Having noted these links between the three papers, I now turn to a separate discussion of each.

I found Makov's review of Bayesian-like approaches to unsupervised sequential learning problems most interesting. This review is especially welcome since most of the work on this topic has appeared outside the mainstream statistical journals. Clearly the problem is a difficult one, and Makov and the others who have worked on various aspects of it are to be congratulated for the progress they have made.

All three cases considered by Makov assume that the p.d.f.'s for an observation x, given that it comes from class H_i, are of the functional form:

$$f_i(x|\theta_i) = f(x|\theta_i,H_i)$$

i.e., there is a common function form for the p.d.f.'s. Moreover, the number of classes, k, is given. The more general problem of mixture with f_i's of possibly different form, and unknown k, has been discussed quite recently by Good and Gaskins (1980 and the ensuing discussion). The computational complexities of the various approaches to more general "bump-hunting" problems make Makov's restrictions quite reasonable for statistical purposes. I should also mention the graphical methods of Fowlkes (1979) for studying mixtures of normals where k is unkown.

I have three questions related to the procedures for Case A discussed in this paper, which may have relatively brief answers:

(1) Has there been any investigation of the adequacy of approximating a mixture of Dirichlets by a single Dirichlet? Good (1967) has noted an example of the inadequacy of a single Dirichlet when the true prior is a mixture.

(2) Isn't part of the problem with the DD, (MDD,) and *PT* methods in your Figure 1 and in other studies due to choice of a "weak" prior?

(3) It would appear that the *QB* computations at step n are not invariant with respect to the ordering of $x_1, \ldots, x_{n-1}$. Is this the case, and if so is it something that a good Bayesian striving for coherence should worry about?

Finally, I note that all of the methods in this paper assumne that observations arrive sequentially, one at a time. Has there been work on related problems when observations arrive in batches?

Smith's paper provides us with a quick, guided tour of the Bayesian approach to change-point problems. It begins in the land of exchangeable subsequences and a consideration of problems representable in such form, and then proceeds with a series of brief stops to explore the more general problems of changes in regression-like structures where the exchangeable subsequences approach is not directly applicable. While the tour has been quick, offering little opportunity for dalliance with any one problem, it has covered much territory in a spirited fashion, and may well whet our appetite for return visits to selected locations.

My comments and queries focus primarily on the simplest of the problems Smith describes in connection with binomial data in Section 2, but I suspect related questions can be raised about the other problems discussed in the later sections. Although the method of analysis described in Section 2 for the Lindisfarne Scribes problem seems quite general, I believe further attention needs to be given to various consistency questions such as the following: (1) If a change-point at $r_1 = 5$ has high posterior probability when $K = 1$ is assumed, does it necessarily follow that $r_i = 5$ will be included in the pair of change-points with the highest posterior probability when $K = 2$? (2) Is it possible that when we place positive prior probabilities on $K = 0,1,2,3$ we can get Bayes factors favoring $K = 2$ at say $(r_1, r_2) = (4,5)$, but when we place positive prior probabilities on $K = 0,1,2,3,4$, we get Bayes factors favoring $K = 4$ at $(r_1, r_2, r_3, r_4) = (4,5,6,7)$? Such consistency properties would seem highly desirable, but would seem to depend on the specification of the priors, $p(\theta_1,\theta_2, \ldots, \theta_{K+1})$. Perhaps Professor Smith has already explored some of these matters in detail.

I also have some concerns regarding the beta structure used for the Lindisfarne Scribes problem. Smith notes that the assignment of independent beta priors may well be unreasonable, but then he goes on to use them nonetheless due to the computational simplicity they provide. Although I have no compelling reasons to suggest in their support, two alternatives that may bear further examination are: (a) a Dirichlet for the joint density of $\theta_j/\Sigma_{i=1}^{k} \ \theta_i, \ j = 1,2, \ldots, K$; or (b) variants of generalized Dirichlets. The major advantage to these densities (aside from the dependencies they introduce) is that it can be represented as a product of independent betas for the random variables $\theta_j/\sum_{i\leq j} \theta_i$. This property may be helpful in achieving the consistency properties I referred to above.

Having revisited the land of exchangeable subsequences with you, I would encourage you to take Smith's complete guided tour for yourselves and choose your own location for an extended visit and prolonged statistical investigation.

While Professor Harrison's oral presentation of this paper can be viewed as nothing short of a tour de force, after several readings of the written version of the paper I am at a loss in my assessment of its contributions to Bayesian decision making. Some of the mathematical aspects of the paper are very interesting, and the discussion of expected loss functions with multiple minima seems quite novel. But there appears to be a fundamental discontinuity in my appreciation and understanding of the paper, as I go from the mathematical formulations to their application. Let me elaborate.

In the initial results they describe, Harrison and Smith investigate decision problems involving bounded utility functions, and they are interested in the behavior of the expected loss, $E[\delta,u]$, with respect to the belief distribution of future outcomes, ϕ, and where δ is the decision, and u represents environmental variables expressible in terms of the parameters of the distribution of ϕ and the loss function, $L(\delta,\phi,u)$. In Section 3, they illustrate that, if the bounded loss function is a monotonically increasing function of δ-ϕ and the distribution of future outcomes is unimodal, then

$E[\delta,u]$ may have two minima. The keys to this result are two: (1) a parameter, α, that appears only in L, (2) monotonic behavior of the scores function over an interval linked to kinks in L. When α takes on one set of values we get the first Bayes decision, and when it takes a different set of values we get a different optimal decision. Despite the fact that slight variations in α may lead to different decision, we must recognize that different values of α do correspond to *different* loss functions, even though they have the same general shape. If one of the two Bayes decisions has smaller expected loss, the decision maker, who after all determines his own loss function, might do well to alter his value of α to achieve the minimum. If, as in the case of the exponential distribution there is a value of α leading to an interval of Bayes decisions with the same expected loss, it does not matter which δ the decision maker chooses since the expected loss does not change. Still, this basic result of Harrison and Smith is somewhat disquieting and bears further examination.

Once we move to multimodal densities for beliefs and for multimodal loss functions, it is not quite so surprising that the possibility of multiple Bayes decisions exists. The main examples Harrison and Smith use to illustrate such situations involve belief distributions which are mixtures of normals and loss functions which are mixtures of Lindley's conjugate normal loss functions. The simplest example here is what the authors refer to as the $E_1^*[\delta]$ model, and involves just the normal distribution with its conjugate loss function, where one of the parameters of the loss function, k, increases with δ. That this situation leads to bifurcating Bayes behavior simply heightens the latent suspicion I have long harboured regarding the appropriateness of Lindley's conjugate loss functions. Yet this result, like the earlier one, is somewhat disquieting.

Up to this point my comments have been technical ones, and have focussed on the mathematical developments described in the paper. The catastrophic discontinuity comes when we turn to the "applications" of this theory.

The first major application of the theory is via an E_2^* model for prison riots. I have been involved in a study of prison-related rehabilitation activities in the United States, have visited with various corrections officials, and actually have spent a little time in a major corrections facility. Thus, I was especially keen to see how a catastrophy-theory like Bayes decision model could be use in "illuminating the dynamics of prison riots"

Now prisons are complex institutions, and to think that an accurate portrayal of behavior in prisons can be made by looking at three crudely defined and artificially interpreted variables seems naive at best. The model *assumed* by Harrison and Smith involves a single normal belief distribution and a bimodal mixture of two conjugate loss functions. Why did they pick such a model? We are told in such loose, heuristic, and ambiguous terms that even the statistically uneducated reader might scream: Stop! What I find even more distressing in the material presented is that we are shown no attempts at model criticism or parameter estimation, the key features of statistical inference when models are used as part of the scientific method (see the related discussion of the role of models in the paper by George Box, given at this conference). You may think from a reading of Section 7.2 that aspects of the requisite data analysis (Bayesian or otherwise) are contained in the referenced papers of Zeeman, Hall, Harrison, Marriage, and Shapland (1976, 1977) but this simply is not so. Although these papers do contain a more detailed description of the data plotted in Figure 7.2, the motivation and justification for the model and the assessment of the adequacy of the model's fit to the data are attended to in a manner just as facile as in the present paper. Even with such a loose approach as the authors choose to present, the model shifts in midstream as a result of "a higher tolerance of tension in the institution after the first mass protest"!

Next we come to the interpretation of the E_2^* model in prison disturbance context. I would claim that all anyone can get out of the model "applied to the data" is what the authors have put into it to begin with. The catastrophes in the prison behavior they "account for" are really only a restatement of the fact that the model contains discontinuities (see the related critique of applied catastrophe theory in the behavioral sciences by Sussman and Zahler, 1978a, 1978b). The Bayesian Bloodhound referred to in Section 7.3, after reading such a description of statistical modelling, would clearly accept this torment no longer, and would visciously attack the authors until they completed a more satisfactory job of analysis and modelling. All of this is not to say that the E_2^* model is inappropriate for the prison riot example (although I have my suspicions). Rather I believe that the authors have not presented very effective evidence in support of their claims.

Finally, to illustrate my concerns with the other major application described in Section 8, I will give my own "application", henceforth to be known as: "Bayesians, Luggage, and the Butterfly Catastrophe". The decision maker involved is a recently-married Bayesian statistician who upon arrival at the Valencia Airport with his wife en route to this Meeting discovers that their luggage has not arrived with them. The loss function involved is much like the one in Section 8.2. There is (1) "pressure from above" by his wife to stay at the airport until the luggage arrives, (2) "pressure from below" by all the other participants who are waiting for the statistician and his wife aboard the bus that is to take them to the Meeting, (3) the statistician's own criterion and beliefs which fall somewhere in the middle. Since this structure is essentially the same as in Section 8.2.1 it should be clear to all readers that we are faced with an example of a butterfly catastrophe. The two conflicting extreme Bayesian decisions can be translated into (1) staying overnight at a hotel near the airport waiting for the luggage to arrive, and (2) immediate departure on the bus. The "middle course" decision is a little too complex to describe here (it involves a non-exponential waiting time distribution with a heavy tail), but we hope to publish a detailed description at a later time. Needless to say, the butterfly cross-section was very helpful in resolving the conflict in this particular problem, as I will indicate quite shortly.

Does the mathematical phenomenon of a butterfly catastrophe follow from my assumptions in a nontrivial way? Indeed, does it really follow at all? Or is this implementation of the E_3^* model simply the consequence of a vague specification, and some hand-waving (perhaps I should say "wing-flapping")? For this example, I readily admit to a contrived "application" of the models and descriptions in the Harrison and Smith paper. I don't believe the resulting butterfly catastrophe tells us anything of practial value at all. Yet I find my own description not all that much different from that of Section 8 of the Harrison and Smith paper. I believe the value of their models in real applications can only be judged by a more careful statistical treatment than the one we are offered in this paper.

All in all, I found the Harrison and Smith paper both stimulating and highly provocative. I look forward to seeing elaborations of their ideas in the future. Lest it appear that I am finishing my comments on a note of dispair, let me note the "luggage example" described above was factual, and that my use of it in this discussion did have one practical consequence. The beleaguered statistician in question (who will remain anonymous) did in fact decide to board the waiting bus and travel without his luggage to the site of the Meeting in Las Fuentes. My description of his plight in the oral presentation of this discussion inspired me to loan him, along with other apparel, my *t*-shirt with the brilliantly-colored image of the famous Dunk Island (Australia) blue butterfly on its chest. Never let it be said that Bayesian decision theory does not have its useful applications!

J.M. BERNARDO (*Universidad de Valencia*):

The need for approximations in the problem discussed by Mr. Makov is fairly clear to me. However, I would like to know more about the quality of the approximation he proposes. For instance, one could try to estimate the expected distance, in some well specified sense, between the exact Bayes posterior predictive distribution

$$\{ p(x_n \epsilon H_i | x_1,\ldots,x_n),\ i = 1,\ldots,k\}$$

and its quasi-Bayes approximation. Moreover, since the true source of the x_i's is never known it might happen that wrong allocations are piled up thus making convergence to the correct allocation as new observations occur difficult, or perhaps impossible.

P.J. BROWN (*Imperial College, London*):

I should like to amplify a point raised by Professor Fienberg concerning the adequacy of the Dirichlet distribution. The problem with the Dirichlet is that it has a very straightjacketed variance-covariance structure. Indeed it involves virtual independence apart from correlations resulting from normalisation to unity. Elsewhere Brown (1976), I have documented some unfortunate features of the Dirichlet prior. To see that an unsupervised learning situation may have a rather different variance-covariance structure consider the following example. There are $K=3$ populations which are $N(\theta_i,1)$, $i=1,2,3$. Imagine the case where θ_2 and θ_3 are close together and quite distant from θ_1. Then unsupervised learning will quickly and accurately determine π_1 and $\pi_2 + \pi_3$ but π_2 and π_3 will be highly correlated together and will have a low correlation with π_1. In this situation the Dirichlet representation will not be able to reflect these second order properties. Thus although Professor Makov's scheme will result in eventual convergence to π_1, π_2 and π_3 it may be difficult to discern the reliability of one's estimates at any stage. Use of an approximating multivariate normal distribution would get around this problem but would of course involve heavier computation.

A.P. DAWID (*The City University, London*):

I want to stress the need for care in setting up models of the kind that Makov has been working with. I am sure these are appropriate for the engineering applications with which he is concerned, but I am none too happy when I see them used in other fields. In particular, I am extremely doubtful about their general suitability in the setting of medical diagnosis, as in the work of Titterington and others.

In this context, the classes $\{H_i\}$ represent diseases, and the observation x corresponds to medical symptoms: thus π may be thought of as describing the prevalence of disease, and θ the "clinical pictures" of the diseases. My unease stems from the seeming possibility, when using a mixture model as described, of obtaining consistent estimates of the parameters. This mean that, if we collect a large enough data-base of patients and record *only* their symptoms (never discovering what diseases they are suffering from), we can nevertheless gain accurate knowledge of both disease prevalences and clinical pictures. My possibly naive reaction to this remarkable state of affairs is one of distrust: how can one learn about anything other than the marginal distribution of symptoms from data such as these? If our model says that we can, it may be a signal that we should discard the model.

It is easy to set up alternative models which behave more reasonably. Instead of splitting up the joint distribution of disease D and symptoms S as $f(d,s|\psi) = f(d|\pi)f(s|d,\theta)$, as is normally done, decompose it instead as $f(d,s|\psi) = f(s|\alpha)f(d|s,\beta)$. (There are good practical reasons for regarding this as more meaningful in a diagnostic context: see Dawid, 1976). We are just as much at liberty to make assumptions about the new parameters (α,β) as about (π,θ). In particular, it does not seem unreasonable to me to assume that α and β are, *a priori*, independent. If so, it is easy to show that observation of patients' symptoms alone will modify the distribution of α, but leave that of β unchanged. Now for purposes of diagnosis (prediction of D from S) only β is relevant: consequently (and, I consider, quite reasonably) such a data-bank is entirely valueless.

A simple example may make my point. Suppose D takes only two values, d_1 and d_2, and likewise S takes values s_1 or s_2. Let $\pi_i = P(D = d_i)$, $\theta_{ij} = P(S=s_j|D=d_i)$, and suppose that π and θ are *a priori* independent, with $\pi_1 \sim \beta(a_{1\cdot},a_{2\cdot})$, $\theta_{11} \sim \beta(a_{11},a_{12})$, $\theta_{21} \sim \beta(a_{21},a_{22})$ all independently (where $a_{i\cdot}=a_{i1}+a_{i2}$). This is an example of Case C of Makov's classification, although it does not satisfy the condition that mixtures should be identifiable. In fact, for this model, a data-bank of unconfirmed cases is quite useless for diagnosis. For, letting $\alpha_j=P(S=s_j)(=\pi_1\theta_{1j}+\pi_2\theta_{2j})$, $\beta_{ij} = P(D=d_i|S=s_j)$ $(=\theta_{ij}\pi_i/\alpha_j)$, we have $P(d,s|\psi) = P(s|\alpha)\ P(d|s,\beta)$, where it is easily found that α and β are *a priori* independent, with, in fact, $\alpha_1 \sim \beta(a_{\cdot 1},a_{\cdot 2})$, $\beta_{11} \sim \beta(a_{11},a_{21})$ $\beta_{12} \sim \beta(a_{12},a_{22})$, all independently. So the considerations of the last paragraph apply.

The moral of all this is that, when we choose a particular mathematical model to represent a real-world process, and make seemingly harmless assumptions (such as "identifiability of mixtures"), we must be careful that any deductions we make are "*qualitatively stable*" in the sense that similar conclusions would be derived from other reasonable ways of modelling the process. (The term 'reasonable" here depends, of course, on the particular application).

Much as I appreciate Makov's contributions to signal detection, I fear that his models may come to be used all too uncritically in other applications, for which they fail to be qualitatively stable.

J.M. DICKEY (*University College of Wales*):

A comment may be of some interest here relative to Professor Smith's paper and other papers in this conference in which an inference is made between nested sampling models. The Bayes factor, or ratio of posterior odds to prior odds in favour of one sampling model versus another, depends on each prior distribution conditional on a model. But what prior distributions should one use and how should they be related? The obvious answer is, of course, that pair of tractable distributions which most closely models actual prior uncertainty conditional on each of the models. Savage's *condition continuity* offers a reasonable guide to such a choice (Dickey and Lientz 1970, Gunel and Dickey 1974). This requires that the prior distribution within the smaller nested model be identical to the conditional distribution induced in the usual way from the joint prior distribution in the larger model. (In this case, the Bayes factor will equal Savage's density ratio, the ratio of posterior to prior densities of the conditioning constraint parameter at the null value).

To see that Professor Smith's choices do not satisfy condition continuity, consider the joint density of the regression coefficient θ and the variance σ^2. The pair (θ,σ^2) are *dependent* under the larger model; the smaller model is obtained by a constraint on θ; hence one will not satisfy condition continuity by having the prior distributions of σ^2 identical under the two models. Professor Smith takes them identical. On the other hand, my own papers on Bayes factors for the normal linear model use condition continuity (Dickey 1971, eq. (5.40), 1974 Prop. 4.2).

Note that by the Borel-Kolmogorov nonuniqueness mentioned in my discussion to Professor Hill's paper in these Proceedings, the answer in each case to the question of whether condition continuity is satisfied will depend on the choice of conditioning constraint variable in terms of which the question is framed. If it is not satisfied for a given pair of distributions for one choice of conditioning variable, perhaps it will be for another choice of conditioning variable. (The same smaller model can often be defined by various essentially different constraints in the larger model). In fact, we shall see this happen for the generalizations of Jeffreys' Bayes factors to be presented later in these Proceedings by Professor Zellner. If the new parameter is used, $\eta = \sigma^{-1}\theta$, and if η is independent of σ, the condition on η will not have an effect on the distribution of σ.

A theorem can even be stated showing that an *arbitrary* given distribution in the smaller model can be obtained by condition continuity from the larger model by suitable choice of conditioning variable. This then would seem to make condition continuity a mathematically vacuous requirement. However, I should like to point out that in practice there are often *natural* conditioning variables η, that is, variables for which one would like to define the smaller model as the consequence of additional information of the form, "η lies in some small hyperinterval centred at the point η_o". This is often the case when the overall mixed type distribution, having positive prior probability attached to the smaller model, is intended as an approximation to a continuous density on the full parameter space with a high mound or ridge over a neighborhood surrounding the constraint set.

In practice, one must be careful to model real uncertainty, and not let the mathematics do one's thinking for one. Professor Smith very wisely took his prior parameters in the binomial Lindisfarne scribe problem so that the uncertainty concerning a single scribe alone was the same as the uncertainty concerning the first scribe among many scribes, instead of the same as if he had been told the many were one. If he had used condition continuity based on any linear conditioning variable, such as $\eta = (\theta_2 - \theta_1, \ldots, \theta_{K+1} - \theta_1)$, then the conditional distribution of θ_1 given $\eta = 0$ would have been beta with parameters $\alpha_1 + \ldots. + \alpha_{K+1} - K$ and $\beta_1 + \ldots + \beta_{K+1}-K$. For $\alpha_i < 1$ and $\beta_i < 1$, $i = 1, \ldots, K+1$, and K large, this distribution of θ_1 would have had a small variance, instead of the same variance as θ_1 under the larger model. That is, if one were told that only one scribe was involved, one's opinion would have been less vague than one's opinion concerning the first scribe of many. Of course, if $\alpha_i < 1$ and $\beta_i < 1$ for all i, then the conditional opinion would have been more, rather than less vague than the unconditional opinion. In fact, for small enough (positive) α_i and β_i, the conditional distribution would have been degenerate, even though the joint distribution was proper. (I am greteful to Professor P.R. Freeman and Professor A.P. Dawid for personal discussions on this example).

J.B. KADANE (*Carnegie-Mellon University*):

The assignment of prior distributions under the different models is a very sensitive matter for model selection. There is no particular justification for the independent beta prior of equation (6), for example. Allowing dependent priors can change the answers in the direction of more scribes, and in fact, can in principle suggest up to 13 scribes. Thus Smith's conclusion that there were probably 3 scribes is heavily dependent on the "perhaps unsatisfactory" assumption (6). The same type of comment applies to the ARMA regression and examples. I welcome, however, the interesting applied problems in this paper, especially the Kidney transplant data.

A more decision-theoretic, and I believe more satisfying approach is developed by Lindley (1968) and continued by Kadane and Dickey (1980).

T. LEONARD (*University of Warwick*):

I think that Professor Smith's approach to change-point inference is very useful and interesting, but I wonder whether he might be over specialising his model simply to facilitate a particular type of conclusion? It would seem to be more natural to assume a Kalman-type model permittting different process levels at each time-stage (e.g. the Harrison-Stevens steady model). A whole range of posterior conclusions could then be reached to suit the practical situation at hand. In particular, we could find the restricted Bayes estimates, for the process levels, amongst a suitable class of step functions, thus providing a very simple way of detecting change-points. Further restricted Bayes procedures would cope with the more complicated situations discussed by Professor Smith. This seems to me to provide a conceptually and technically simple way of coping with change-points, and avoids choosing an unduly complex model simply to cope with a single very special type of posterior conclusion. These aspects have been discussed by Leonard (1978).

REPLY TO THE DISCUSSION

U.E. MAKOV (*Chelsea College, London*):

I wish to thank our discussants for their interesting questions and comments and make the following points:

1.- The choice of a Dirichlet prior in *case A* was made simply to exploit the conjugacy property. As described in the paper, neither the 'unfortunate properties of the Dirichlet prior' nor the possible inadequacy of approximating a mixture of Dirichlets by a single Dirichlet affect the desirable assymptotic properties of the QB procedure. However, these inadequacies are bound to affect the small sample properties of the procedure and I suspect that in acute situations, like the one suggested by Dr. Brown, the QB might be painfully slow.

One possible remedy, which was tried numerically,. is to approximate a mixture of size n by a mixture of size k, $k < n$. In Makov (1978) the mixture of Dirichlets is allowed to grow so long as is computationally possible (rather than collapsing the mixture after *each* observation). Thereafter, the mixture build-up can be restarted. In Makov (1978), the mixture, once collapsed, was replaced by a single Dirichlet and no further growth was allowed. In Smith and Makov (1980), in the context of *case B*, the approximating mixture consists of Gaussian p.d.f's. Here, in the case of detection and estimation of jumps in linear systems, the combined quality of the detector/estimator is considerably improved when the number of terms in the approximating density is increased.

2.- The QB procedure is not invariant to the order of the observations, an undesirable property in small samples. One possible solution is to take the observations in batches, where each batch is processed coherently (Bayes) and then approximated by a QB procedure. (In Smith and Makov (1980) batches of 2 and 4 observations were used). Another possibility is to choose two (or more) possible sequences (the most likely in some sense) and then to average the QB estimators for these sequences.

3.- The consistency of the QB recursion (when so proved), as opposed to the possible asymptotic bias of the DD and PT, is inherited in the mathematical structure of the recursion and is invariant to the choice of prior. According to Stochastic-Approximation theory, certain properties of the regression $E\{\pi\text{-}w_1\}$ (see (18) above) are required for consistency. The analysis of such regressions shows that consistency is not affected by the choice of priors but by the degree of overlap between the densities of the corresponding classes. While the QB remains consistent for any degree of overlap (though its rate of convergence is affected), the other methods are consistent only for overlaps below a certain threshhold (or signal-to-noise ratio larger than some value). This also explains why the QB is asymptotically immune to initial errors (or wrong allocations).

4.- The only QB qualities investigated were asymptotic unbiasness and relative efficiency. We have no 'distance measure' to compare the QB with the coherent Bayes procedure.

5.- Prof. Dawid expresses doubts about the possibility of obtaining consistent estimator for both the 'clinical picture' and 'prevalences of disease'. His reaction cannot be contradicted as no proof of such consistency exists for *case C*, to which he refers. As for *case A*, and several models in *case B*, such consistency is proved on the basis of identifiability. When this assumption cannot be made, the QB should not be adopted, nor should any other procedure which is based on an inappropriate model!

There are, however, cases in the medical context where the assumption of identifiability is acceptable. For instance (see Hermans and Habbema (1975)), in the diagnosis of Haemophilia carriership the identifiable mixture consists of two bivariate normal densities whose means and covariances are estimated from the data, while the mixing parameters are established through genetical considerations. Though the QB may prove to be consistent for this problem, I have my own reservations about the adequacy of its use (and indeed of the exploitation of uncorfirmed cases as a whole) in the case of small samples. (See Makov (1980) for further details).

I am not at all sure how qualitatively stable is the choice of *linear discriminant* function in medical diagnosis. However, in recent papers*, (O'Neill, 1978; Ganesalingam and McLachlan, 1978), it is shown that the ratio of the relevant (*asymptotic*) information contained in unclassified observation to that of classified observation is quite considerable for a statistically interesting range of separation of the populations. In Ganesalingam and Mclachlan (1979), simulation studies of *small* misclassified samples produced satisfactory results.

A.F.M. SMITH (*University of Nottingham*):

The points which have concerned the discussants of my paper also concern me.

(i) I agree with Professor Kadane that model selection criteria should really be derived using an appropriate loss or utility framework. My current work on these problems is now proceeding along decision-theoretic lines.

(ii) While the consistency properties mentioned by Professor Feinberg might seem appealing at first sight, it is not clear to me that they would be implied by all specifications of priors: I have not succeeded (yet) in sorting out precisely when they would hold.

(iii) The general points raised by Professor Dickey concerning "condition continuity" are very interesting and have been well-aired at this conference. As he himself admits, however, there is often considerable arbitrariness in the way in which a smaller model is derived from a larger by conditioning and this leaves us with the problem of providing a rationale for any particular choice. Dickey is, of course, correct in noting that my regression specification violates condition continuity, but I am also using the improper form $p(\sigma) \propto \sigma^{-1}$ and suspect that when *this* pragmatic approximation doesn't make me feel too uncomfortable neither will I feel too bad about the other.

(iv) Dr. Leonard may well be correct in suggesting that other formulations of the change-point problem could lead to simpler ways of detecting change; I look forward to seeing further details.

* I am indebted to Dr. D.M. Titterington for these references.

P.J. HARRISON (*Warwick University*):

We wish to thank Professor Fienberg for his amusing comments and hope to clarify some of the points which seem to have been misunderstood. We shall begin by dealing with the theoretical points.

(i) It is *not* the kinks in our loss function L which generate the pertinent discontinuities. Smooth loss functions with no jumps can exhibit the same kind of discontinuous trajectory of the corresponding Bayes decision. It is simply *easiest* to illustrate this behaviour by using a double step loss function which just happens to be discontinuous.

(ii) Although different values of α give different loss functions it should be noted that any loss function combines a utility with a function representing quantifiable loss (see DeGroot, 1970). It would be a brave man who would suggest that he knew this utility function *precisely* or, indeed, that it did not change with the decision-maker's environment. However, we have shown that slightly different utilities can give rise to extremely different Bayes decisions even when the posterior density is from a smooth and well-known family.

(iii) We hold that bounded loss functions should always be used in a Bayesian analysis. The discontinuities we discuss here cannot be considere a *fault* of using the normal conjugate loss function. Indeed, if we use quadratic loss, for example, then the corresponding expected loss function does not represent the decision-maker's dilemma when he is faced with a very bimodal posterior density. This indicates to us that this form of analysis must be lacking in some fundamental way. Under practical considerations bounded loss is a necessity due to the boundedness of the resources of the decision-maker. Theoretically the use of unbounded loss $L(\delta\text{-}\theta)$ can give rise to some very awkward paradoxes. For example if $L(\delta\text{-}\theta)$ is convex then it is easily shown that the corresponding Bayes decisions always depend solely on the comparative steepness at $\mp\infty$ of the two tails of a posterior density on the real line (see Kadane and Chuang, 1978). We must therefore reconcile ourselves to the fact that under any sensible analysis it is possible to get these sudden changes in decision.

We shall now reply to the discussion about some of the examples that we presented. Obviously in the time and space available we were only able to give the bare bones of the structure of the analysis of what are vast macro-models. A little more detail will be presented in Smith, Harrison and Zeeman (1979) and Smith (1980) has developed the ideas given in Zeeman et al (1974). Because of this lack of space we considered it most important to indicate how discontinuous phenomena can be analysed in a Bayesian way so that the underlying discontinuities of the system are not obliterated by our model. It is obvious, but not always realised, that before we can *criticize* a statistical model we must choose one. Either we acknowledge the absence of developed theory, and construct our own model, or we undertake the often very difficult task of translating a theoretical model (e.g. chemical or psychological) into an appropriate statistical one. For the prison riot case study we had to translate the theories of Konrad Lorentz first into a mathematical and then a Bayesian model. This translation was informative in itself, necessitating, for example, the use of *bounded* loss structures in our description. Since Lorentz's model is qualitative it generates a *class* of statistical models containing posterior densities and loss functions geometrically "similar" to the normal and double conjugate respectively. Any sensibly parametrised model in this class will give the same kind of geometry to the data. We worked with the normal and its conjugate solely for computational ease. Professor Fienberg's contention that the reader can get nothing out of the model other than the data is quite untrue since the model is now in a statistical form and therefore the parameters can be estimated. These estimates together with other information can be combined by any self-respecting Bayesian to give predictive distributions for the outbreaks of violence as functions of 'Alienation" and "Tension". These are in fact being used in some British institutions. However this methodology, being long and technically dull, would have been out of place amongst the contributions presented at Valencia.

It seems very common, in forecasting and other fields, for a statistician to construct a model with no regard to the dynamics of the underlying process and just "fitting" it. Consequently the practitioner has little information communicated to him other than a short term forecast which he probably could have achieved by eye anyway. We sincerely hope that Professor Fienberg is not proposing this as the ideal (and only) function of a statistician. Although we enjoyed the paper presented by Professor Box at this conference, we felt that he might have emphasized the importance of picking a sensible class of models to begin with. Unfortunately it seems likely that his paper may be used by some statisticians as an excuse to avoid essential thought.

REFERENCES IN THE DISCUSSION

BROWN, P.J. (1976). Remarks on some statistical methods for medical diagnosis. *J. Roy. Stat. Soc.* **A 139**, 104-107.

DAWID, A.P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647-658.

DICKEY, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.* **42**, 204-223.

— (1974). Bayesian alternatives to the *F* test and the least squares estimate in the normal linear model. In *Studies in Bayesian Econometrics and Statistics* (S.E. Fienberg and A. Zellner, eds.) 515-554. Amsterdam: North Holland.

DICKEY, J.M. and LIENTZ, B.P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *Ann. Math. Statist.* **41**, 214-26.

FOWLKES, E. (1979). Some methods for studying the mixture of two normal (log-normal) distributions. *J. Amer. Statist. Assoc.* **74**, 561-575.

GANESALINGAM, S. and MCLACHLAN, G.J. (1978). The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika* **65**, 658-662.

— (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. *J. Statist. Comput. Simul.* **9**, 151-158.

GOOD, I.J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. B* **29**, 399-431.

GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite Data. *J. Amer. Statist. Assoc.* **75**, 42-73.

GUNEL, E. and DICKEY, J.M. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545-57.

HERMANS, J. and HABBEMA, J.D.F. (1975). Comparison of five methods to estimate posterior probabilities. *EDV in Medizin and Biologie,* 14-19.

KADANE, J.B. and DICKEY, J.M. (1980). Bayesian Decision Theory and the Simplification of Models. To appear in *Criteria for Evaluation of Econometric Models,* (J. Kmenta and J. Ramsey, eds.)

KADANE, J.B. and CHUANG, O.T. (1978). S table decision problems. *Ann. Statist.* **6**, 1095-1110.

LEONARD, T. (1978). Density Estimation, Stochastic Processes, and Prior Information (with discussion). *J. Roy. Statist. Soc. B* **40**, 113-146.

LINDLEY, D.V. (1968). The Choice of Variables in Multiple Regression (with discussion). *J. Roy Statist. Soc. B* **30**, 31-66.

MAKOV, U.E. (1978). An algorithm for sequential unsupervised classification. *Proceedings in Computational Statistics, "Comsptat 1978"* (Corstan, L.C.A. and Hermans, J., eds.).

— (1980). The statistical problem of unconfirmed cases in medicine. To appear in *Teoria delle Decisioni in Medicina.* (E. Girelli-Bruni ed.). Verona: Bertani.

O'NEILL, T.J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.* **73**, 821-826.

SMITH, A.F.M. and MAKOV, U.E. (1980). Bayesian detection and estimation of jumps in linear systems. *Proceedings of the IMA Conference on "The Analysis and Optimization of Stochastic Systems".* (O.R.L. Jacobs *et al.* eds.), 333-346. New York: Academic Press.

SMITH, J.Q. (1980). The Prediction of Prison Riot. *J. Math. Statist. Psychol.* (To appear).

SUSSMAN, H.J. and ZAHLER, R.S. (1978a). A critique of applied catastrophe theory in the behavioral sciences. *Behavioral Science,* **23**, 383-389.

— (1978b), Catastrophe theory as applied to the social and biological sciences: A critique. *Synthese* **37**, 117-216.

Trab. Estadist. (1980)
Vol. **31** pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 141-204 Valencia, Spain: University Press

3. Likelihood, sufficiency and ancillarity

INVITED PAPERS

AKAIKE, H. (*The Institute of Mathematical Statistics, Tokyo*)
Likelihood and the Bayes procedure.

DAWID, A.P. (*The City University, London*)
A Bayesian look at nuisance parameters.

DISCUSSANTS

BARNARD, G.A. (*University of Waterloo, Canadá*)
FREEMAN, P.R. (*Leicester University*)
PEÑA, D. (*Escuela de Organización Industrial, Madrid*)
DICKEY, J.M. (*University College Wales*)
GEISSER, S. (*University of Minnesota*)
LINDLEY, D.V. (*University College London*)
O'HAGAN, A. (*University of Warwick*)
SMITH, A.F.M. (*University of Nottingham*)

REPLY TO THE DISCUSSION

Likelihood and the Bayes procedure

HIROTUGU AKAIKE

The Institute of Statistical Mathematics, Tokyo

SUMMARY

In this paper the likelihood function is considered to be the primary source of the objectivity of a Bayesian method. The necessity of using the expected behavior of the likelihood function for the choice of the prior distribution is emphasized. Numerical examples, including seasonal adjustment of time series, are given to illustrate the practical utility of the common-sense approach to Bayesian statistics proposed in this paper.

Keywords: LIKELIHOOD; BAYES PROCEDURE; AIC; SEASONAL ADJUSTMENT.

1. INTRODUCTION

The view that the Bayesian approach to statistical inference is useful, practically as well as conceptually, is now widely accepted. Nevertheless we must also accept the fact that there still remain some conceptual confusions about the Bayes procedure. Although many strong impetuses for the use of the procedure came from the subjective theory of probability, it seems that the confusions are also caused by the subjective interpretation of the procedure.

By looking through the works on the Bayes procedure by subjectivists, it quickly becomes clear that there is not much discussion of the concept of likelihood. The subjective theory of probability is used only to justify the use of the prior distribution of the parameters of a data distribution. It is almost trivial to see that no practically useful Bayes procedure is defined without the use of the likelihood function, while the likelihood function can be defined without the prior distribution. Thus the data distribution represents the basic part of our prior information and the Bayes procedure gives only one specific way of utilizing the information supplied by data through the likelihood function.

From this point of view there is nothing special about the choice of prior distributions to differentiate it from the design of ordinary statistical procedures such as the choice of the sampling procedure in a sample survey and the choice of the spectrum window in the spectrum analysis of a time series.

In this paper we first discuss some conceptual confusions with the Bayes procedure which we believe to be due to the subjective interpretation of the procedure. We argue that it is necessary to recognize the limitation of the subjective theory and put more emphasis on the concept of likelihood. We take the position of regarding the Bayes procedure as one possible way of utilizing the information provided by the likelihood function. Once such an attitude towards the Bayes procedure is accepted we can freely develop Bayesian models simply by representing a particular preference of the parameters by a prior distribution. The goodness of the prior distribution can then be checked by evaluating expected performances of the corresponding Bayes procedure in various conceivable situations.

We demonstrate the use of this type of approach by developing a general Bayesian model for the analysis of linear relations between variables. The model contains as special cases the basic models of those estimation procedures such as the Stein estimator, ridge regression, Shiller's distributes lag estimator and O'Hagan's localized regression. Numerical examples are given to illustrate the practical utility of some quasi-Bayesian procedures developed for these models and for a more conventional model of polynomial regression. The result of application to the seasonal adjustment of time series seems particularly interesting as the model contains twice as many parameters as the number of the observations.

2. CONCEPTUAL DIFFICULTIES OF THE SUBJECTIVE APPROACH

Significant impetus for the advancement of Bayesian statistics has come from the side of the subjective theory of probability. This is natural as every statistical procedure may be viewed as a formulation of the psychological process of information processing and evaluation by a skilful researcher. In spite of the significant contribution of the subjective theory of probability to clarifying the nature of the psychological aspect of this process, several conceptual difficulties remain with the theory. Here we discuss some difficulties, which we believe to be misconceptions, related to the Bayes procedure and clear the way for the development of practically useful Bayesian methods.

2.1. *Rationality and Savage's axiom*

It is sometimes said that a rational person must behave as if he has a clearly defined system of subjective probabilities of uncertain events. This is often ascribed to Savage (1954) who developed a theory of personal probability by axiomatizing the preference behavior of a person under uncertainty. Unfortunately the very first postulate P1 of Savage, which assumes the linear ordering of the preference, excludes the real difficulty of preference. This can be explained by the following simple example.

Consider a young boy who wants to choose a girl as his wife. His preference is based on the three characteristics, H, I and L. Here H stands for health, I for intelligence and L for looks. Each characteristic is ranked by the numbers 1, 2, and 3, with higher number denoting higher rank. The difference of ranks by 1 is marginal and the difference by 2 means a significant difference. Denote by $R_i = (H_i, I_i, L_i)$ the vector of the ranks of the characteristics of the i^{th} girl. Being uncertain about the relative importance of these characteristics in his future life, he ignores the marginal differences and pays attention only to the significant differences. Thus his preference is defined by the following scheme:

$$R_i \leq R_j, \text{ i.e., the j}^{\text{th}} \text{ girl is preferred to the i}^{\text{th}} \text{ girl,}$$

$$\text{iff } C_i \leq C_j \text{ for the characteristic } C$$

$$\text{for which } |C_j - C_i| \text{ is maximum.}$$

Now he has three girl friends (i = 1, 2, 3) whose R_i's are defined by $R_1 = (1, 2, 3)$, $R_2 = (3, 1, 2)$ and $R_3 = (2, 3, 1)$, respectively. Obviously it holds that

$$R_1 \leq R_2, R_2 \leq R_3 \text{ and } R_3 \leq R_1,$$

which shows that his natural preference scheme does not satisfy the postulate P1 of Savage.

It is the difficulty of this type of preference that make us feel the need of a horoscope or some other help in making the decision in a real life situation. Since Savage's system excludes the possibility of this type of difficulty, the corresponding theory of personal probability cannot tell how we should treat the difficulty. The exact characterization of Savage's theory is then a theory of one particular aspect of preference and there is no compelling reason to demand that a rational person's preference should be represented by a single system of subjective probability. Wolfowitz (1962) presents a pertinent discussion of this point. Thus to justify the use of a system of personal probability one must prove its adequacy by some means. Certainly the proof cannot be found within the particular system of personal probability itself.

2.2. *The role of parameters in a Bayesian modeling*

The subjective theory of probability of De Finetti demands that the probability distribution or the expectations of the uncertain events of interest should completely be specified (de Finetti, 1974b, p. 87). If we accept this demand and decide to use the Bayes procedure, all we have to do is to compute $p(y|x)$, the probability of an event y conditional on a given set of data x. The theory only asserts that the necessary probability distribution should be there, and does not consider the special role played by the parameters in constructing a statistical model or the probability distribution. De Finetti (1974a, p. 125) even rejects the concept of a parameter as metaphysical, unless it is a decidable event.

That the concept of parameter cannot be eliminated is shown by the simple example of the binomial experiment where the probability of occurence of a head in a coin tossing is considered. The concept of independent trials with a fixed probability of head is unacceptable by the subjective theory of probability of de Finetti and the solution is sought in the concept of exchangeability (de Finetti, 1975, pp. 211-218). The difficulty is caused by the fact that the probability of a head, which must be decided, plays the role of a parameter that is not actually decidable (Akaike, 1979b).

We may use the theory of probability to develop some understanding of what we psychologically expect of the parameters of a statistical model. Consider a random variable x and the observations x_1, x_2, ...of some related events. We expect that a parameter θ exhausts the information about x to be gained through the observations x_1, x_2... The probabilistic expression of this expectation is given by

$$p(x|\theta, x_1, x_2, ...) = p(x|\theta), \qquad (2.1)$$

where $p(x|z_1, z_2, ...)$ denotes the distribution of x conditional on z_1, z_2, ... To allow this type of discussion we must consider θ as a random variable as is advocated by Kudo (1973). The formula (2.1) then gives a very natural characterization of the parameters as a condensed representation of the information contained in the observations, i.e., once θ is known no further observations can improve our predictions on x. Thus we want to know the value of θ. Actually de Finetti's discussion of the exchangeable distribution of the binomial experiment has given a proof of the existence of such a variable.

Although the above characterization of a parameter is interesting, in the statistical model building for inference the order of reasoning is reserved. The prior information first suggests what type of parameterization of the data distribution $p(x|\theta)$ should be used. The prior distribution $\pi(\theta)$, if at all specified, represents only a part of the prior information. To take the parameters as something prespecified and assume that the prior distribution can or should be determined independently of the data distribution constitutes a serious misconception about the inferencial use of the Bayes procedure.

2.3. *Likelihood principle and the Bayes procedure*

It has often been claimed that the likelihood principle, which demands that the statistical inference should be identical if the likelihood function is identical, is a direct consequence of the Bayesian approach; see, for example, Savage (1962, p. 17). In the example of coin tossing, if we denote the probability of head by θ and assume the independence and homogeneity of the tossings, we have

$$p(x|\theta) = C\, \theta^x (1 - \theta)^{n-x}$$

as the likelihood of θ when x heads appeared in n tossings. It is argued that there is no difference in the inference through the Bayes procedure if the above likelihood is obtained as the result of n tosses, with n predetermined, or as the result of tossing continued until x heads appeared, with x predetermined.

This seemingly innocuous argument is against the principle of rationality of the subjective theory of probability which suggests that the choice of a statistical decision be based on its expected utility. The expected behavior of the likelihood function $p(x|\theta)$ is certainly different for the two schemes of the coin tossing and it is irrational to adopt one and the same prior distribution $\pi(\theta)$, irrespectively of the expected difference of the statistical behavior of the likelihood functions.

To clarify the nature of the confusion by a concrete example, consider the use of the posterior distribution

$$C\, \theta^x(1 - \theta)^{n-x}\, \pi(\theta)$$

as an estimate of the probability distribution of the result y of the next toss, where $y = 1$ for head and 0 otherwise. The predictive distributions are defined as the averages of the data distribution $p(y|\theta)$ with respect to the posterior distributions of θ. These will be denoted by $p(y|x)$ and $p(y|n)$ to indicate that x and n are the realizations of the random variables, respectively. They are defined by

$$p(y|*) = C \int_0^1 \theta^{x+y}(1 - \theta)^{n+1-x-y}\pi(\theta)d\theta,$$

where $*$ stands for either x or n. When the "true" value of θ is θ_o the goodness of $p(y|*)$ as an estimate of the true distribution $p(y|\theta_o) = \theta_o^y (1 - \theta_o)^{1-y}$ can be measured by the entropy of $p(y|\theta_o)$ with respect to $p(y|x)$ or $p(y|n)$ which is defined by

$$\mathrm{B}\,\{p(\cdot|\theta_o), p(\cdot|*)\}$$

$$= -\Sigma_{y=0}^{1}\left\{\frac{p(y|\theta_o)}{p(y|*)}\right\}\log\left\{\frac{p(y|\theta_o)}{p(y|*)}\right\}p(y|*)$$

The larger the entropy the better is the approximation of $p(\cdot|*)$ to $p(\cdot|\theta_o)$. Before we observe x or n we evaluate $\mathrm{E}_*\,\mathrm{B}\{p(\cdot|\theta_o), p(\cdot|*)\}$ for some possible values of θ_o, where E_* denotes the expectation with respect to the distribution of $*$ defined with $\theta = \theta_o$. We have

$$\mathrm{E}_x\,\mathrm{B}\{p(\cdot|\theta_o), p(\cdot|x)\}$$

$$= \Sigma_{y=0}^{1} p(y|\theta_o)\,\Sigma_{x=0}^{n}\,\log\left\{\frac{p(y|x)}{p(y|\theta_o)}\right\}{}_nC_x\,\theta_o^x\,(1-\theta_o)^{n-x}$$

and

$$\mathrm{E}_n\,\mathrm{B}\{p(\cdot|\theta_o), p(\cdot|n)\}$$

$$= \Sigma_{y=0}^{1}\,p(y|\theta_o)\,\Sigma_{n=x}^{\infty}\,\log\left\{\frac{p(y|n)}{p(y|\theta_o)}\right\}{}_{n-1}C_{x-1}\,\theta_o^x\,(1-\theta_o)^{n-x}.$$

Obviously we have no reason to expect that these two quantities will take one and the same value and, at least for that matter, there is no reason for us to assume one and the same prior distribution $\pi(\theta)$ for both cases.

3. LIKELIHOOD AS THE SOURCE OF OBJECTIVITY

The discussion in the preceding section illustrates both the subjective and objective elements in the Bayesian approach to statistical inference. It is subjective because a statistical inference procedure is designed to satisfy a subjectively chosen objective. The choice of the data distribution is particularly subjective and the prior distribution reflects the object of the inference which is often expressed in the form of a psychological expectation.

What is then objective with the procedure ? The objectivity stems from the dependence on the data which is a production of the outside world. This objectivity is fed into the Bayes procedure through the likelihood function. Since $\mathrm{B}\{p_o(\cdot), p(\cdot|\theta)\} = \mathrm{E}_x \log p(x|\theta) - \mathrm{E}_x \log p_o(x)$, we can see that, ignoring the additive constant $\mathrm{E}_x \log p_o(x)$, the log likelihood $\log p(x|\theta)$ is a natural estimate of the entropy of $p_o(\cdot)$ with respect to $p(\cdot|\theta)$. Here E_x denotes the expectation with respect to the distribution $p_o(\cdot)$ of x. Thus the likelihood $p(x|\theta)$ represents an objective measure of the goodness, as measured by x, of $p(\cdot|\theta)$ as an approximation to $p_o(\cdot)$. This fact forms the basis of the practical utility of the Bayes procedure even for the family $\{p(\cdot|\theta)\}$ which is chosen subjectively and does not contain the true distribution of x.

The likelihood function $p(x|\theta)$ is the basic device for the extraction or condensation of the information supplied by the data x. The role of the prior distribution $\pi(\theta)$ is to aid further condensation of the information supplied by the likelihood function $p(x|\theta)$ through the introduction of some particular preference of the parameters. By evaluating the expected entropy of the true distribution with respect to the predictive distribution specified by a posterior distribution we can extend the concepts of bias and variance to the posterior distribution (Akaike, 1978a). If we try to keep a balance between the bias and variance, we cannot ignore the influence of the statistical behavior of the likelihood function on the choice of our prior distribution. Some of the conflicts between the conventional and Bayesian statistics are caused by ignoring the possible dependence of the choice of the prior distribution, or even the choice of the basic data distribution, on the number of available observations which influences the behaviour of the likelihood function; see, for example, Lindley (1957), Schwarz (1978) and Akaike (1978b).

4. A GENERAL BAYESIAN MODELING FOR LINEAR PROBLEMS

In this section we demonstrate the practical utility of the point of view discussed in the preceding section through the discussion of a general Bayesian model for the analysis of linear problems. The basic idea here may be characterized as the common-sense approach to Bayesian statistics.

Consider the analysis of the linear relation between the vector of observations $y = [y(1),\ldots,y(N)]'$ and the vectors of the independent variables $x_i = [x_i(1), x_i(2),\ldots x_i(N)]'$ $(i = 1,2,..,K)$, where $'$ denotes transposition. The method of least squares leads to the minimization of

$$L(a) = \Sigma_{j=1}^{N}\,[y(j) - \Sigma_{i=1}^{K} a_i x_i(j)]^2. \tag{4.1}$$

We know, when K is large compared with N or when the matrix $X = [x_1, x_2,\ldots,x_K]$ is ill-conditioned the least squares estimates behave badly. To control this we introduce some preference on the values of the parameters and try to minimize

$$L(a) + \mu\|a - a_0\|_R^2 \tag{4.2}$$

where a_0 denotes a particular vector of parameters $[a_{01}, a_{02},\ldots,a_{0K}]'$, $\|\ \|_R^2$ the norm defined by a positive definite matrix R, and μ a positive constant. The use of this type of constrained least squares for the solution of an ill-posed problem is wellknown; see, for example, Tihonov (1965).

The difficulty with the application of this method of constrained least squares is in the choice of the value of μ. To solve this we transform the problem into the maximization of

$$\ell(a) = \exp\left\{-(1/2\sigma^2)\,[L(a) + \mu\|a - a_0\|_R^2]\right\},$$

where temporarily σ^2 is assumed to be known. Since we have

$$\ell(a) = \exp[-(1/2\sigma^2)\,L(a)]\,\exp[-(\mu/2\sigma^2)\,\|a - a_0\|_R^2],$$

we can see that the solution of the constrained least squares problem is now given as the mean of the posterior distribution defined by the data distribution

$$f(y|\sigma^2, a) = (1/2\pi)^{N/2}(1/\sigma)^N \exp[-(1/2\sigma^2)\,L(a)], \tag{4.3}$$

and the prior distribution

$$\pi(a|d) = (1/2\pi)^{K/2}(1/\sigma)^K \exp[-(d^2/2\sigma^2)\,\|a - a_0\|_R^2], \tag{4.4}$$

where $d^2 = \mu$. By properly choosing X, a_0 and R, we can get many practically useful models. Particularly, we will restrict our attention to the case where $\|a - a_0\|_R^2$ is defined by

$$\|a - a\|_R^2 = \|c_0 - Da\|^2, \tag{4.5}$$

where D is a properly chosen matrix, $c_0 = Da_0$ and $\|v\|^2$ denotes the sum of squares of the components of v. In this case the posterior mean of the vector parameter a is obtained by minimizing $\|z(a|d)\|^2$ of the vector $z(a|d)$ defined by

$$z(a|d) = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \\ dc_0(1) \\ dc_0(2) \\ \vdots \\ dc_0(L) \end{bmatrix} - \begin{bmatrix} X \\ \\ aD \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \vdots \\ \\ a(K) \end{bmatrix} \tag{4.6}$$

Examples.

a. *Stein type shrunken estimator*

This is defined by putting $L = N$, $D = X$ and $c_o = 0$, the zero vector. The case with $K = N$ and $X = I_{N\times N}$ corresponds to the original problem of estimation of the mean vector of a multivariate Gaussian distribution treated by Stein. By putting c_o equal to the vector of the parameters obtained from some similar former observations, we can realize a reasonable use of the prior information.

b. *Ridge regression*

This is defined by putting $L = K$, $D = I_{K\times K}$ and $c_o = 0$.

c. *Shiller's distributed lag estimator*

Shiller (1973) developed a procedure for the estimation of a smoothly changing impulse response sequence. In this case $[y(1), y(2), \ldots, y(N)]$ is-obtained as the time series of the output of a constant linear system under the input $u(j)$. X is defined by $x_i(j) = u(j-i+1)$ and $c_o = 0$.
D is put equal to

$$D_1 = \begin{bmatrix} \alpha & & & & & \\ -1 & 1 & & & 0 & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & 0 & & \ddots & \ddots & \\ & & & & -1 & 1 \end{bmatrix}$$

or

$$D_2 = \begin{bmatrix} \alpha & & & & & & & \\ -\beta & \beta & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & 0 & & \\ & & \ddots & \ddots & \ddots & & & \\ & 0 & & & & & & \\ & & & & & 1 & -2 & 1 \end{bmatrix}$$

where α and β are properly chosen constants. D_1 controls the first order differences of $a(j)$ and D_2 the second order differences.

d. *Localized regression of O'Hagan.*

O'Hagan (1978) introduced an interesting Bayesian model for the estimation of the locally gradually changing regression of a time series $y(i)$ on $x(i)$. Our model corresponding to O'Hagan's is given by putting $K = N$, $c_o = 0$ and

$$X = \begin{bmatrix} x(1) & & & 0 \\ & x(2) & & \\ & & \ddots & \\ 0 & & & x(N) \end{bmatrix}.$$

D is put equal to D_1 or D_2 of the above example or

$$D_3 = \begin{bmatrix} \alpha & & & & & & & & \\ -\beta & \beta & & & & & & & \\ \gamma & -2\gamma & \gamma & & & & 0 & & \\ -1 & 3 & -3 & 1 & & & & & \\ & -1 & 3 & -3 & 1 & & & & \\ & & \ddots & \ddots & \ddots & \ddots & & & \\ & 0 & & & & & & & \\ & & & & & -1 & 3 & -3 & 1 \end{bmatrix}.$$

One particularly interesting model is obtained by putting $x(i) = 1$ $(i = 1,2,\ldots, N)$. The number of parameters in this model is equal to the number of observations $y(i)$.

e. *Locally smooth trend fitting*

For a time series $y(i)$, by putting $c_o = 0$ and $D = D_k X$ where D_k is as given in the preceding examples, we get a model for the fitting of a smooth trend curve. One special choice of X is given by $X = I_{NxN}$. We will call the model defined with $X = I_{NxN}$ and $D = D_k$ the model of locally smooth trend of k^{th} order.

f. *Bayesian seasonal adjustment*

We consider the decomposition of the monthly observations $y(i)$ for M years, where $i = 12m + j$ $(j = 1,2, \ldots, 12,\ m = 0,1, \ldots, M\text{-}1)$, into the form

$$y(i) = T_i + S_i + I_i,$$

where T_i denotes the trend, S_i the seasonal and I_i the irregular component. For this problem we put $K = 2N$ $(N = 12M)$ and define $a = (T_1, T_2, \ldots, T_N, S_1, S_2, \ldots, S_N)$ and put $c_o = 0$.

The matrix X is defined by

$$X = \begin{array}{c|c} & \xleftrightarrow{\quad N \quad} \qquad\qquad \xleftrightarrow{\quad N \quad} \\ \updownarrow N & \begin{bmatrix} 1 & & & & & 1 & & & & \\ & 1 & & & & & 1 & & & \\ & & 1 & 0 & & & & 1 & 0 & \\ & & & \ddots & & & & & \ddots & \\ & 0 & & & & & 0 & & & \\ & & & & 1 & & & & & 1 \end{bmatrix} \end{array}$$

and D by

$$
D_{kp} = \begin{array}{c|c:cccccc|}
 & \leftarrow N \rightarrow & \multicolumn{6}{c}{\longleftarrow N \longrightarrow} \\
\uparrow N \downarrow & D_k & & & 0 & & & \\
\hdashline
\uparrow & & eI & 0 & 0 & & & 0 \\
 & & fI & -fI & 0 & & & 0 \\
N & 0 & 0 & fI & -fI & & & 0 \\
 & & \vdots & \vdots & & \ddots & \ddots & \vdots \\
\downarrow & & 0 & 0 & & & fI & -fI \\
\hdashline
\uparrow & & g1' & 0 & & & & 0 \\
 & & 0 & g1' & & & & 0 \\
N & 0 & \vdots & \vdots & & \ddots & & \vdots \\
\downarrow & & 0 & 0 & & & & g1' \\
 & \leftarrow N \rightarrow & \leftarrow 12 \rightarrow & \leftarrow 12 \rightarrow & & & \multicolumn{2}{c}{\longleftarrow 12 \longrightarrow}
\end{array}
$$

where D_k is one of those defined in the preceding examples, $I = I_{12\times 12}$, $1' = (1, 1, \ldots, 1)$, and e, f, g are properly chosen constants.

A notable characteristic of this model is that it has twice as many parameters as the number of observations. This constitutes a typical many parameters problem which cannot be handled by the ordinary unconstrained least squares or the method of maximum likelihood.

The fundamental problem in applying these models to real data is the choice of the constant d. Assuming that other constants are specified, the decision on d is equivalent to the decision on the prior distribution of a. From (4.2) the choice of d, or μ, determines the relative weight of the additional term $\| a - a_o \|_R^2$ against $L(a)$, the sum of squares of the residuals. When a_o is not exactly equal to the true value of a, we expect that the bias of the estimate increases as d is increased but the variance decreases. It is natural to try to keep a balance between these two factors. To realize this it is necessary not to specify d uniquely but use the information supplied by the likelihood function or $L(a)$.

In the Bayesian terminology this is to consider d as a hyperparameter which has its own prior distribution. Now it is obvious that by considering d as a hyperparameter we are trying to use the information supplied by the likelihood function for the determination of d. This observation suggests that a proper choice of the prior distribution to be used in an inferential situation can only be realized through the analysis of the statistical characteristics of the related likelihood function. The infinite digression of considering the priors of priors can only be stopped by the analysis of the expected output at each stage, which is determined by the behavior of the likelihood function.

Incidentally, the present observation shows why the conventional subjectivist doctrine of assuming the determination of the prior distribution of the parameters independently of the related likelihood function was not strictly followed by the research workers dealing with real inference problems. This point is discussed as the Bayes / Non-Bayes compromise by Good (1965). We take here the very flexible attitude towards the Bayes procedure to consider it only as one possibility of utilizing the information supplied by the likelihood function. Thus we consider that any practically useful statistical procedure which utilizes the information supplied by the likelihood function should not be rejected only because it is non-Bayesian. It is not the dogmatic exclusion of other procedures but the explicit proposal of useful models that proves the advantage of the Bayesian approach over the conventional statistics.

5. NUMERICAL EXAMPLES

To show that our discussion in the preceding sections is not vacuous, here we show some numerical examples. These were obtained by Bayesian modelings but with the help of some procedures which are not strictly Bayesian. The first three examples are concerned with the models discussed in the preceding section. The last one is an example of polynomial fitting and is included to show the feasibility of a Bayesian modeling with the aid of an information criterion (AIC) to deal with the difficulty of choosing a prior distribution for a multimodel situation where the models are with different number of parameters.

For the first three examples the essential statistic used for the determination of the parameter d in (4.4) is the likelihood of the model specified by the prior distribution. We consider the marginal likelihood of (d,σ^2) defined by

$$L(d,\sigma^2) = \int f(y\,|\,\sigma^2,a)\ \pi\,(a\,|\,d)\ da,$$

where $f(y\,|\,\sigma^2,a)$ and $\pi(a\,|\,d)$ are given by (4.3) and (4.4), respectively. If we assume (4.5) and put $c_o = 0$ we get

$$L(d,\sigma^2) = (1/2\pi)^{N/2}(1/\sigma)^N \exp\,[\,-(\,1/2\sigma^2)\,\|\, z(a_*\,|\,d)\,\|^2]$$

$$\cdot \,\|\, d^2D'D \,\|^{1/2} \,\|\, d^2D'D + X'X \,\|^{-1/2},$$

where $\| z(a_*|d) \|^2$ denotes the minimun of $\| z(a|d) \|^2$ with $z(a|d)$ defined by (4.6). Instead of developing a prior distribution of (d,σ^2) we consider the use of the procedure which chooses a model with the maximum marginal likelihood. This is called the method of type II maximum likelihood by Good (1965). For a given d, the maximum with respect to σ^2 is attained at

$$\sigma_d^2 = (1/N)\| z(a_*|d)\|^2.$$

For the case of practical applications, we consider a finite set of possible values $(d_1, d_2, \ldots, d_l)$ of d and choose the one that maximizes $L(d,\sigma_d^2)$. Since we are familiar with the use of minus twice the log likelihood, we propose to minimize

$$\text{ABIC} = (-2) \log L(d, \sigma_d^2)$$

$$= N \log [\, 1/N \| z(a_*|d) \|^2 \,] + \log \| d^2D'D + X'X \|$$

$$- \log \| d^2D'D \| + \text{const},$$

where ABIC stands for "a Bayesian information criterion". When different D's are not considered, the term $\log \| d^2D'D \|$ may be replaced by $2K \log d$, where K is the dimension of the vector a.

In the last example we demonstrate the practical utility of $\exp(-\frac{1}{2}\text{AIC})$ as the definition of the likelihood of a model specified by the maximum likelihood estimates of the parameters. Here AIC is by definition (Akaike, 1974)

$$\text{AIC} = (-2) \log \text{(maximum likelihood)} + 2 \text{ (number of free parameters)}.$$

This definition allows a very practical procedure of developing a Bayesian type approach to the situation where several models with different numbers of parameters are considered.

The general definition of ABIC of a model with hyperparameters determined by the method of type II maximum likelihood would have been ABIC = (-2) log (maximum marginal likelihood) + 2 (number of adjusted hyperparameters). In the examples treated in this paper the numbers of the adjusted hyperparameters are identical within the models being compared and their influence on the maximum marginal likelihoods is ignored.

Examples

a. *Distributed lag estimation*

We did a simulation with the second example of Shiller (1973, p. 783). The result is illustrated in Table 1. This result was obtained by using the model

c of the precedin section with $N = 40$, $K = 20$ and $D = D_2$ with $\alpha = \beta = 0$. Considering that this is a limiting situation with non-zero α and β, ABIC was defined by

$$\text{ABIC} = N \log [\, (1/N) \| z(a_*|d)\|^2]$$

$$+ \log \| d^2D'D + X'X \| - 2\,K \log d,$$

and the ABIC was minimized over $d = 5.0, 2.5, 1.25, 0.625, 0.3125$. the values of the ABIC at these d's were -43.4, -51.5, -52.7, -45.0, -30.9, respectively. The minimum, -52.7, was attained at $d = 1.25$ and corresponding estimates of the parameters are given in Table 1 along with the theoretical values and the least squares estimates. By taking a properly weighted average of the results with different d's we may get a procedure which has smaller sampling variability, but it seems that the present simple procedure is almost sufficient for many practical applications.

TABLA 1
Example of distributed lag estimation

	1	2	3	4	5
Theoretical	.000	.000	.001	.004	.018
Bayes	-.009	-.003	.004	.009	.017
Least squares	-.010	.021	-.045	.037	.078
	6	7	8	9	10
Theoretical	.054	.130	.242	.352	.399
Bayes	-.051	.134	.242	.345	.395
Least squares	-.074	.255	.113	.462	.334
	11	12	13	14	15
Theoretical	.352	.242	.130	.054	.018
Bayes	.362	.257	.134	.052	.012
Least squares	.359	.329	.046	.072	.042
	16	17	18	19	20
Theoretical	.004	.001	.000	.000	.000
Bayes	-.001	-.015	.006	.035	-.018
Least squares	-.018	-.050	.065	-.008	-.008

b. *Locally smooth trend fitting*

In this example the original data $y(i)$ $(i = 1, 2, \ldots, 30)$ were generated by the relation

$$y(i) = 4 \exp [- (1/2) ((i-5)/4)^2] + z(i),$$

where z(i)'s are independently and identically distributed as $N(0,1)$. Twelve models of locally smooth trend of k^{th} order defined by the model e of the preceding section with $d = 2^{8-j}$ ($j = 1, 2, \dots, 12$) were tried with $k = 1, 2, 3$. The constants α, β, and γ of the D_k's were all put equal to 0.001. The ABIC was defined by

$$\text{ABIC} = N \log [(1/N) \| z(a_* | \text{d}) \|^2] + \log \| \text{d}^2 D'D + X'X \|$$

$$-\log \| d^2 D'D \|.$$

The minimum of ABIC was attained at $k = 1$ and $d = 2.0$. The original data, the theoretical trend and some of the estimated trends are illustrated in Fig. 1. In this figure SSDEV stands for the sum of squares of deviations of the estimates from the theoretical. It can be seen that the present procedure can produce meaningful results even with these rather noisy observations. In the figures ID stands for k.

c. *Seasonal adjustment*

In this case the model f was applied to various artificial and real time series of length six years, i.e., $N = 72$. The constants of D_k in D_{kp} were the same as in the preceding example and other constants were $e = 0.001$, $f = 1.0$ and $g = 10.0$. The set of twelve values of d used in the preceding example was also used here and $k = 1, 2, 3$ were tried. Results corresponding to the minima of the ABIC's are illustrated in Fig.'s 2—4.

Fig. 2 shows the result of application of the present procedure to an artificial series given in Abe, Ito, Maruyama et al (1971, pp. 250-251). The result shows a very good reproduction of the true trend curve which was disturbed by a fixed multiplicative seasonality and the addition of the irregular components to produce the observations denoted by original.

It is remarkable that by this procedure no special treatment is necessary at the end of the series. This point is a significant advantage over the conventional procedures which require various ad hoc adjustments at the beginning and end of the series (Shiskin and Eisenpress, 1957). Fig. 3 shows the result of application to the last six years of the series of the logarithms of the number of airline passengers, given as Series G in Box and Jenkins (1970). The result reveals a very reasonable gradual change of the seasonality. The procedure has also been applied to the time series of labor force given in Table 1 of Shiskin and Eisenpress (1957, p.442) and the result is given in Fig. 4. The adjusted series is simply defined by y(i) - S, and is compared with the series adjusted by the Method II by Shiskin and Eisenpress.

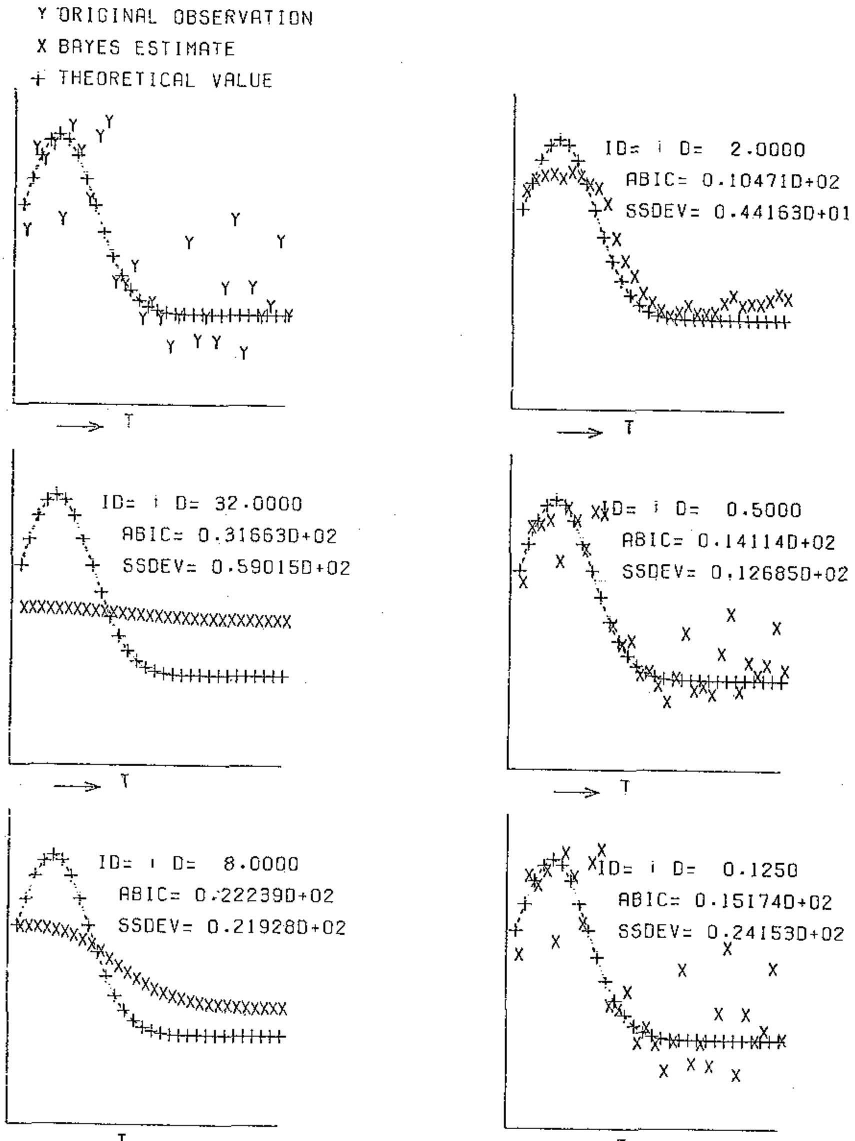
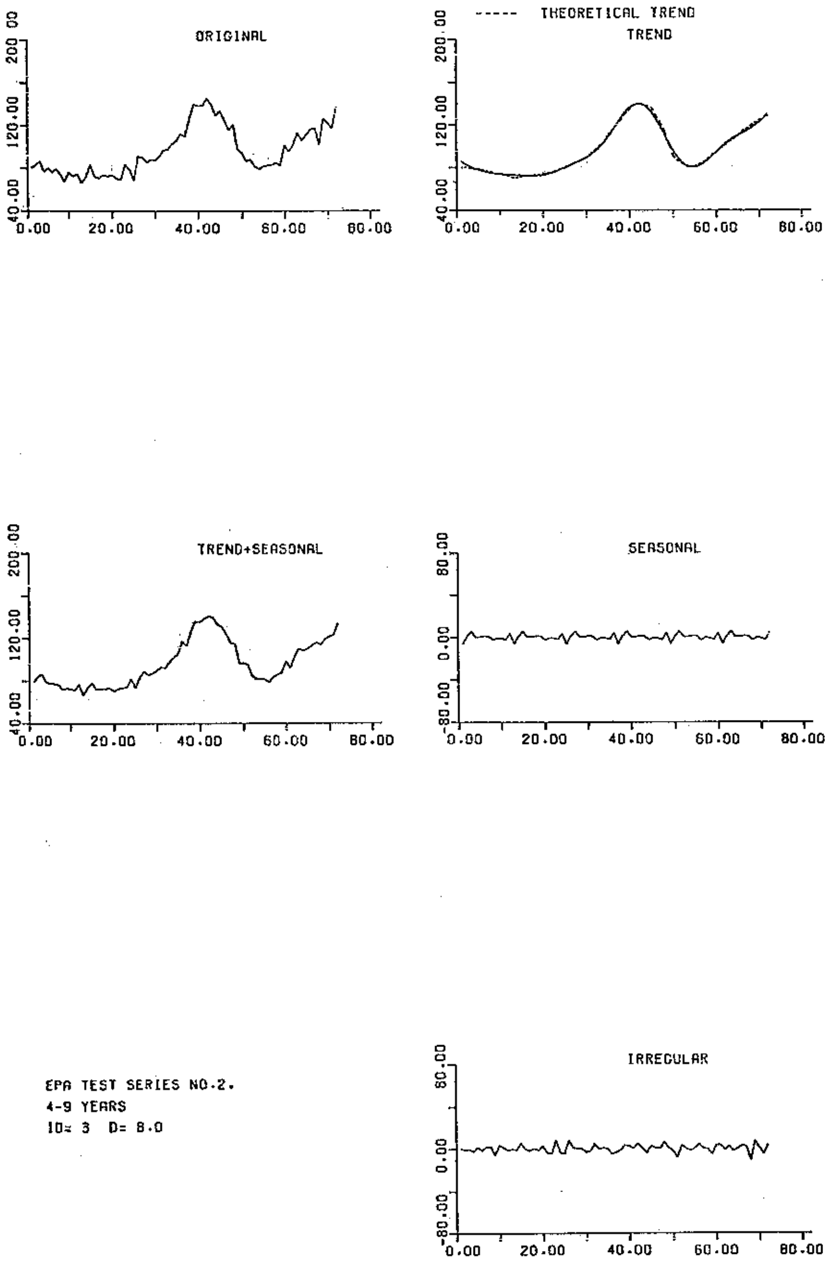
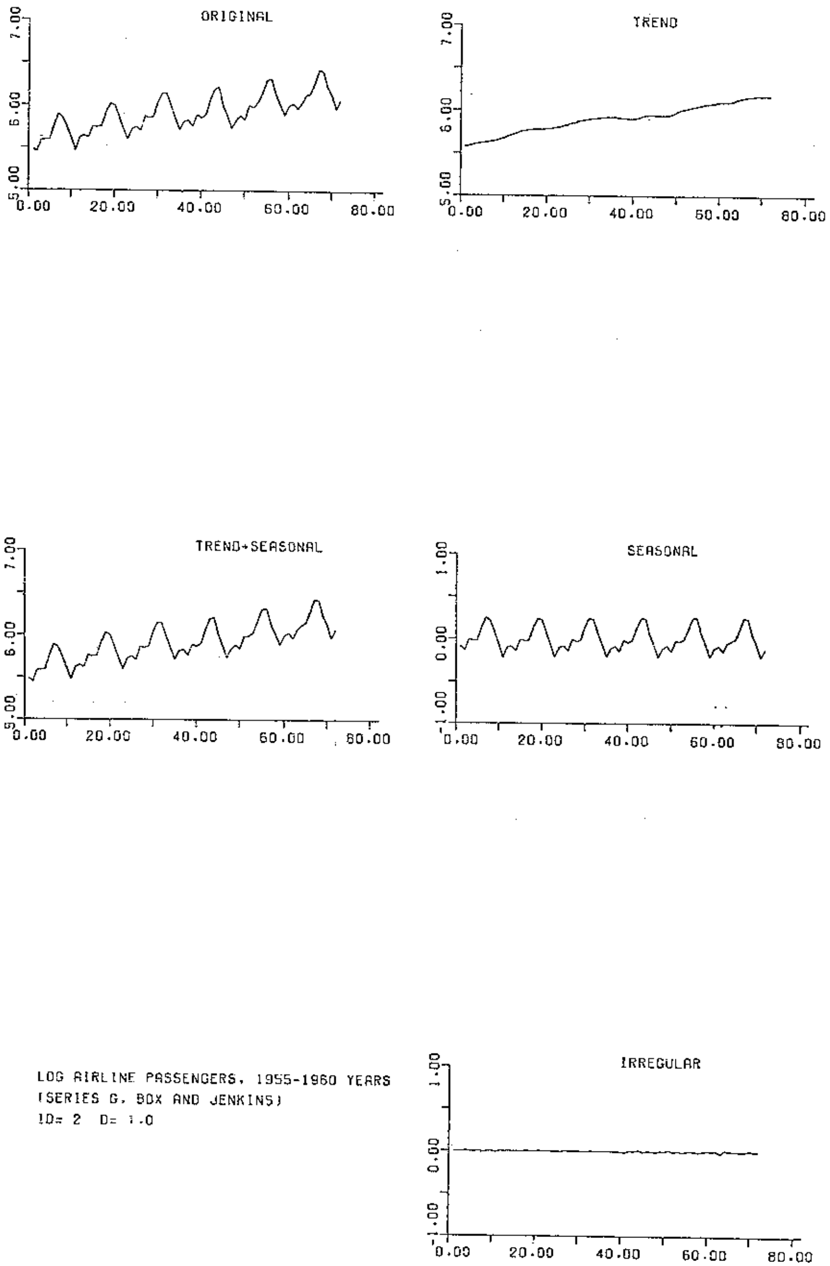


FIGURE 1



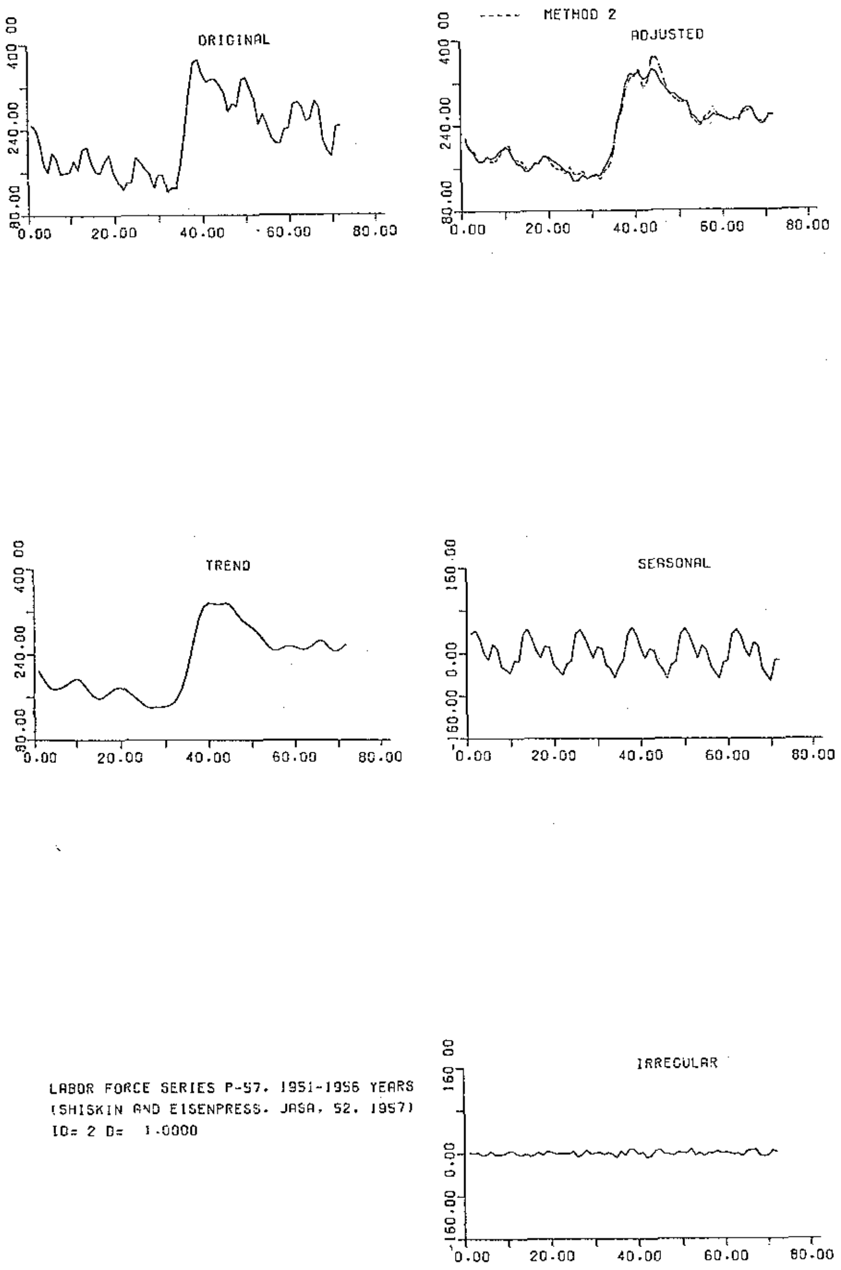
EPA TEST SERIES NO.2.
4-9 YEARS
ID= 3 D= 8.0

FIGURE 2



LOG AIRLINE PASSENGERS, 1955-1960 YEARS
(SERIES G, BOX AND JENKINS)
ID= 2 D= 1.0

FIGURE 3



LABOR FORCE SERIES P-57. 1951-1956 YEARS
(SHISKIN AND EISENPRESS. JASA, 52, 1957)
ID= 2 D= 1.0000

FIGURE 4

d. *Polynomial fitting*

By this example we wish to demonstrate that a reasonable definition of the likelihood of a model defined by the maximum likelihood estimates of the parameters can be given by exp (-(1/2) AIC) (Akaike, 1979a, c). The observations $y(i)$ are identical to those of the example *b* of this section and the polynomials of successively increasing order were fitted up to the 10th order by the method of maximum likelihood. Under the assumption of the Gaussian distribution, the AIC of the M^{th} order model is defined by

$$\mathrm{AIC}\ (M) = N \log \left[\, (1/N)\ \mathrm{S}(M)\, \right] + 2M,$$

where $S(M)$ denotes the sum of squares of the residuals. Some of the estimated regression curves and the values of the AIC are illustrated in Fig. 5.

We smoothed these regression curves with the weight proportional to exp $[\, -\ (1/2)\ \mathrm{AIC}\ (M)\,]\ \pi(M)$ with $\pi(M) \propto (M + 1)^{-i}$. The result is denoted by "Bayes" in the figure. The same type of procedure has been applied to the fitting of autoregressive models by Akaike (1979a) where the choice of $\pi(M)$ is discussed.

The present result shows that the procedure is practically useful, although its performance depends on the choice of the system of the basic functions or the polynomials. Usually this choice produces significant effects at the beginning and end of the regression curve. This shows the advantage of the models used in the preceding examples *b* and *c* over the present model. Nevertheless the present result demonstrates the feasibility of a Bayesian modeling of a multi-model problem with models defined with different number of parameters.

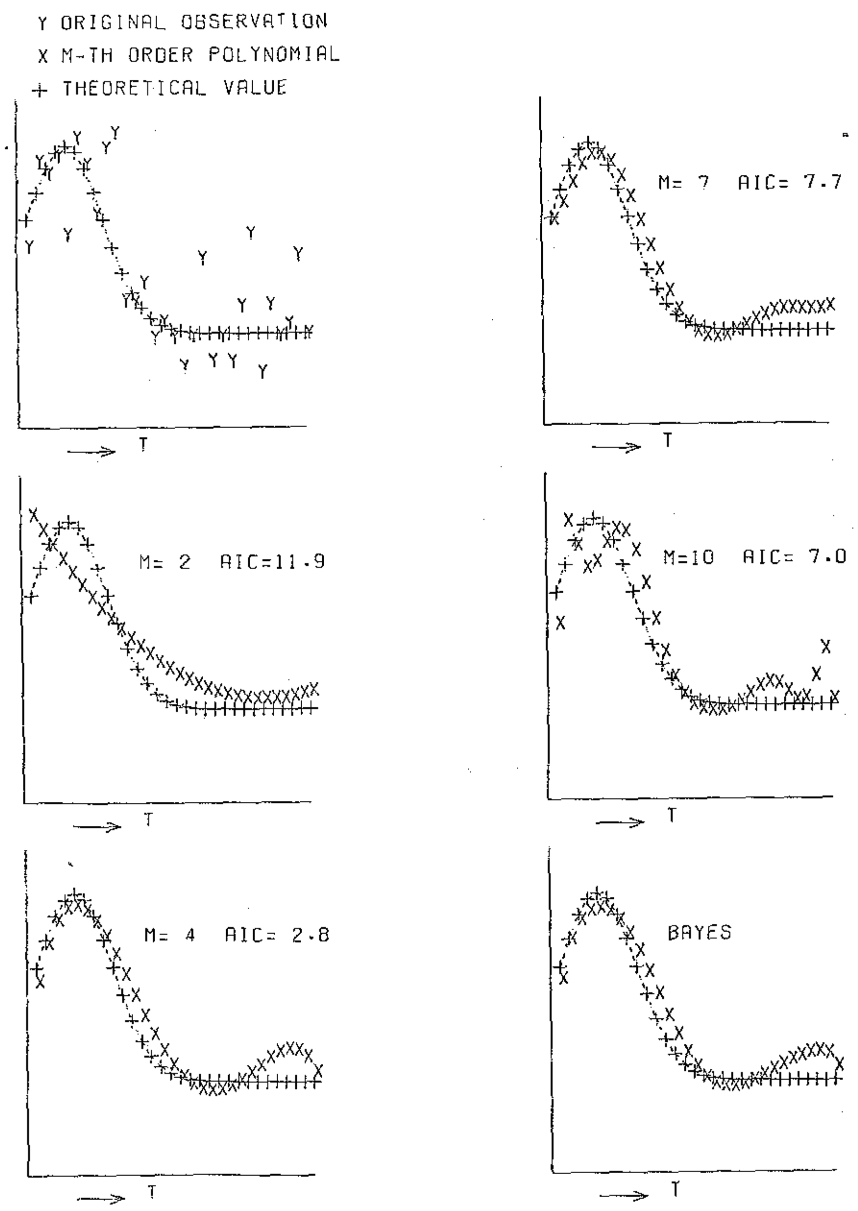


FIGURE 5

6. DISCUSSION

The numerical results presented in the preceding section suggest the possibility of developing further applications of the general linear model to problems such as the gradually changing autoregression and the general trend analysis of time series. This possibility is pursued in Akaike (1979d). By choosing the set of d's properly the type II maximum likelihood method may be replaced by a procedure which takes an average of the models with respect to the weight proportional to the likelihood of each model. The performance of these procedures are controlled by the statistical characteristics of the related likelihood functions. One particular possibility is the extension of the concept of ignorance prior distribution to the prior distribution of a hyperparameter. This is discussed in Akaike (1980).

The application to seasonal adjustement is particularly interesting as it provides an example of the model which cannot be treated by the ordinary method of maximum likelihood. This example clearly demonstrates the practical utility of the Bayesian approach. It also shows that our present procedure may be characterized as a tempered method of maximum likelihood. The practical utility of the general linear model stems from the understandability and manipulability of the related prior distributions. This allows us to make proper judgement on how to temper the likelihood function through the choice of the values of the constants within the priors.

The subjective theory of probability is developed on the basis of our psychological reaction to uncertainty. Acordingly the final justification of the theory must be sought in the psychological satisfaction it can produce throught its application to real problems. It is only the accumulation of successful results of application that can really make the Bayesian statistics attractive.

The Bayes procedure provides a natural and systematic way of utilizing the information supplied by a likelihood function. The likelihood has a clearly defined objective meaning as the measure of the goodness of a model. It is this objectivity that provides the basis for the use of the subjective theory of probability as a guide in developing statistical procedures. Only this objectivity allows us to develop our confidence on the practical utility of the Bayes procedure, even when we know that the related model is our subjective construction.

ACKNOWLEDGEMENTS

The author is grateful to Dr K. Tanabe for the stimulating discussion of the linear model. The author is also indebted to Ms. E. Arahata for preparing the numerical and graphical outputs reported in this paper.

REFERENCES

ABE, K; ITO, M., MARUYAMA, A., YOSHIKAWA, J., ISUKADA, K. and IKEGAMI, M. (1971). *Methods of Seasonal Adjusiements. Research Series No. 22.*, Tokyo: Economic Planning Agency Economic Research Institute (In Japanese).

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control,* **AC-19,** 716-723.

— (1978a). A new look at the Bayes procedure. *Biometrika,* **65,** 53-59.

— (1978b). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.,* **30,** *A,* 9-14.

— (1979a). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika,* **66,** 53-59.

— (1979b). A subjective view of the Bayes procedure. *Research Memo. No. 117.* Tokyo: The Institute of Statistical Mathematics. Revised, February 1979.

— (1979c). On the use of the predictive likelihood of a Gaussian model. *Research Memo. No 159.* Tokyo: The Institute of Statistical Mathematics.

— (1979d). On the construction of composite time series models. *Research Memo. No 161.* Tokyo: The Institute of Statistical Mathematics.

— (1980). Ignorance prior distribution of a hyperparameter and Stein's estimator. *Ann. Inst. Statist. Math.,* **32,** *A,* 171-178.

BOX, G.E.P. and JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control.* San Francisco: Holden Day.

DE FINETTI, B (1974a). Bayesianism: Its unifying role for both the foundation and application of statistics. *Ini. Stat. Rev.,* **42,** 117-130.

— (1974b/1975) *The Theory of Probability, Volumes 1 and 2.* New York: Wiley.

GOOD, I.J. (1965). *The Estimation of Probabilities.* Cambridge, Massachusetts: M.I.T. Press.

KUDO, H. (1973). The duality of parameter and sample. *Proceedings of the Institute of Statistical Mathematics Symposium,* **6,** 9-15 (In Japanese).

LINDLEY, D.V. (1957). A statistical paradox. *Biometrika,* **44,** 187-192.

O'HAGAN, A. (1978). Curve fitting and optimal design for prediction. *J.R. Statist. Soc. B,* **40,** 1-42.

SAVAGE, L.J. (1962). Subjective probability and statistical practice. In *The Foundations of Statistical Inference.* (G.A. Barnard and D.R. Cox eds.) 9-35.London: Methuen.

— (1954) *The Foundations of Statistics.* New York: Wiley.

SCHWARZ, G. (1978) Estimating the dimension of a model. *Ann. Statist.,* **6,** 461-464.

SHILLER, R. (1973) A distributed lag estimator derived from smoothness priors. *Econometrica,* **41,** 775-778.

SHISKIN, J. and EISENPRESS, H. (1957). Seasonal adjustments by electronic computer methods. *J. Amer. Statist. Ass.,* **52,** 415-499.

TIHONOV, A.N. (1965). Incorrect problems of linear algebra and a stable method for their solution. *Soviet Math. Dokl.,* **6,** 988-991.

WOLFOWITZ, J. (1962). Bayesian inference and axioms of consistent decision. *Econometrica,* **30,** 470-479.

A Bayesian Look at Nuisance parameters

A. P. DAWID

The City University, London

SUMMARY

The elimination of nuisance parameters has classically been tackled by various *ad hoc* devices, and has led to a number of attempts to define partial sufficiency and ancillarity. The Bayesian approach is clearly defined. This paper examines some classical procedures in order to see when they can be given a Bayesian justification.

Keywords ANCILLARY; BAYESIAN INFERENCE; D-ANCILLARY; D-SUFFICIENT; G-SUFFICIENT; NUISANCE PARAMETER; S-ANCILLARY; S-SUFFICIENT; SUFFICIENT.

1. INTRODUCTION

Problems with nuisance parameters, where we are interested in only a part of the parameter that governs the distribution of our data, are of prime practical importance, yet our theoretical understanding remains limited and confused. One attractive approach is to simplify the model, by reducing the data or by conditioning on some statistic. Attempts to justify such simplification may be based on generalization of the concepts of sufficiency and ancillarity, but this generalization may be made in many ways. Another approach is to be, or to act like, a Bayesian, and integrate out any unwanted parameters. This in itself leads to a particular form of generalized sufficiency and ancillarity which, while of little direct interest to the Bayesian, is useful as a standard for judging other definitions.

In this paper we use examples and theory to indicate both the similarities and the differences between the Bayesian and classical approaches. Section 2 introduces nuisance parameters. Section 3 and 4 describe the Bayesian approach to generalized sufficiency, and Section 5 some classical definitions. In Section 6 we illustrate various possible ways in which these properties can hold, alone or together. Section 7 introduces specialized Bayesian versions of generalized sufficiency and ancillarity which are of particular relevance for comparison with the classical approach, and Section 8 takes up this comparison. In particular, it is shown that, under certain conditions, the classical approach can be given a Bayesian justification only for very special prior distributions.

Notation.

A capital letter will normally be used to denote an uncertain quantity (random variable or parameter), and the corresponding small letter for a realized or hypothetical value. However, this convention is not rigid. We use the symbols f, $\bar{f}$ and π to denote probability densities, leaving the relevant variables to be understood from the context: thus $f(t|s,\theta)$ denotes the density at t for the distribution of T, conditional on $S = s$, when $\Theta = \theta$. Our manipulations with such densities will be informal and far from rigorous, though all can be made precise. Thus $f(x|\lambda) = f(t|\theta)f(x|t,\phi)$ means that the parameter Λ is equivalent to the pair (Θ,Φ); that the marginal distributions of T depend on Λ through Θ alone; and that the conditional distributions of X given T depend on Φ alone. Such concepts can be conveniently and accurately expressed using the notation of *conditional independence* (Dawid, 1979a): the above properties would read: $T \perp \Lambda \mid \Theta$, $X \perp \Lambda \mid (T,\Phi)$. However, as this notation is still relatively unfamiliar, it has been avoided in this paper.

2. NUISANCE PARAMETERS

Suppose we are interested in the value of some unknown quantity Θ, (which, like all other abstract quantities we shall consider, may have several components) and can conduct a statistical investigation to learn about Θ. The outcome of this investigation will be our data x, the realised value of a random variable X.

If we are fortunate, the distribution of X will be completely determined by the value of Θ; this is the state of affairs treated in greatest depth in the inference text-books. However, in most real problems such simplicity is an unattainable ideal, even after we have made simplifying assumptions, such as normality, in setting up the model. Instead, the distributions of X might be governed by a parameter Λ, which is in some way connected with Θ. The most common case, to which we shall restrict our attention in this paper, is that Θ gives only a partial description of the distributions of X, so that Θ is a function of Λ.

The usual approach to such a problem is to introduce a further parameter Φ which, combined with Θ, completes the specification of the distribution of X. Then the pair (Θ, Φ) may be taken to be Λ. For example, if our experiment consists of an unbiased measurement of Θ, where the measuring instrument is subject to a normally distributed error of unknown variance, it would be usual to take Φ to be this variance. In such a case, Φ would be designated as "the nuisance parameter", and inference about Θ becomes "elimination of Φ". This seems a natural and obvious stance, but it should be pointed out that there is an arbitrariness involved in the choice of nuisance parameter. For instance, in the above case, why not take, for Φ, the *coefficient of variation* of

the distribution? There is, indeed, a whole host of possible choices of the nuisance parameter. For some purposes (in particular, Bayesian inference) this will make no difference to our inference about Θ but, as we shall see, it may frequently be important to recognise the arbitrary nature of the nuisance parameter.

3. THE BAYESIAN APPROACH

A coherent Bayesian B has no conceptual difficulty in making inference about Θ in the presence of nuisance parameters. The distributions of X depend only on Λ, so that the observation $X = x$ provides a likelihood function, $f(x|\lambda)$ say, for Λ. To use this coherently requires a prior distribution for Λ, which B can specify. He now derives, in the usual way, his posterior distribution for Λ and, being interested in Θ alone, simply summarises his posterior opinions about Θ by means of the implied marginal posterior distribution for Θ.

No specification of a nuisance parameter is neccesary for this calculation, and if such a choice is made —as it normally will be— it is for convenience alone. For example, knowledge of the real world problem at hand may often make it possible to choose a Φ for which it would be reasonable to take Θ and Φ as *a priori* independent. In the measurement problem of the previous section, this might pick out the variance, rather than the coefficient of variation; but is easy to think of similar problems, with the same normal family of distributions, where this preference might be reversed. In any event, such a choice of nuisance parameter serves merely to simplify the psychological problem of specifying one's prior distribution, and is in no way essential to the statistical analysis.

In general, for any choice of Φ, B's distribution of Λ can be re-expressed as a joint distribution for (Θ,Φ), which can then be decomposed into the marginal distribution for Θ (which may be easy to specify, and will not depend on which Φ is used) and a conditional distribution for Φ given Θ (which may not, and of course will). Representing parameter-densities by the symbol π, Bayes'theorem gives

$$\begin{aligned} \pi(\theta,\phi|x) &\propto \pi(\theta,\phi) f(x|\theta,\phi) \\ \text{and so} \quad \pi(\theta|x) &\propto \int f(x|\theta,\phi)\pi(\theta,\phi)d\phi \\ &= \int f(x|\theta,\phi)\pi(\theta)\,\pi(\phi|\theta)d\phi \\ &= \bar{f}(x|\theta)\,\pi(\theta) \end{aligned}$$

where $\bar{f}(x|\theta) = \int f(x|\theta,\phi)\ \pi(\phi|\theta)d\phi$ gives the density of B*'s* coherent distributions for X given only that $\Theta = \theta$, which we shall call the *marginal model* (for B). The marginal model does not depend on the choice of nuisance parameter Φ. As a function of θ, $\bar{f}(x|\theta)$ is known as the *marginal likelihood* of Θ, based on data $X = x$.

From the point of view of the single Bayesian B, the marginal likelihood is as good as any ordinary likelihood, but there are differences so far as the whole scientific community is concerned. There will normally be a good measure of agreement about the full model $f(x|\theta,\phi)$ (is not this what we really mean by a model?); but the marginal model $\bar{f}(x|\theta)$ is constructed by an operation involving *B's* subjective opinions, through $\pi(\phi|\theta)$, and so does not appear to share the objectivity of $f(x|\theta,\phi)$.

Armed with the marginal model, we can consider such concepts as sufficiency and ancillarity in it. We shall call T *marginally sufficient* for Θ(for B), if it is sufficient in the marginal model, and similarly for marginal ancillarity. Note that these concepts depend on the prior distribution, but only through the *conditional* distributions for Λ given Θ, the marginal prior distribution for Θ being arbitrary. Thus a collection $\boldsymbol{B}$ of Bayesians, with various prior distributions $\{\Pi_B : B \in \boldsymbol{B}\}$ for Λ, will all agree on the marginal model, and so agree whether or not a statistic T is marginally sufficient or ancillary, as long as they agree on the model and on the distributions of Λ given Θ (in which case we shall call $\boldsymbol{B}$ a *bevy* of Bayesians). An alternative statement of this last condition is that, for the family $\{\Pi_B\}$ of distributions, regarded as a model with "data" Λ and "parameter" B, Θ is a sufficient "statistic".

4. MARGINAL SUFFICIENCY

Suppose T is marginally sufficient for Θ (for B). Then $\bar{f}(x|\theta)$ has the form $a(x)\,\bar{f}(t|\theta)$ where $\bar{f}(t|\theta)$ is the marginal density of T given $\Theta = \theta$. Thus $\pi(\theta|x) \propto \pi(\theta)\,\bar{f}(x|\theta) \propto \pi(\theta)\,\bar{f}(t|\theta)$, whence $\pi(\theta|x) = \pi(\theta|t)$, and *B's* posterior marginal distribution for Θ depends on T alone, just as in the case of ordinary sufficiency with no nuisance parameters. Under some regularity conditions, the converse will hold. Our definition is therefore in accord with those of Raiffa and Schlaifer (1961) and Lindley (1965).

In a sense, marginal sufficiency is unimportant: B will get the same posterior distribution for Θ whether he bases it on the complete data X or on T alone, and for this very reason there is little point in his reducing his data to T before processing. However, it is often necessary to discard some data in the interests of manageability, and if B knows that he can do this in such a way that he loses no information about Θ, so much the better.

This raises the question: How can B know that T is marginally sufficient? It seems that he must first either evaluate his posterior for Θ, and discover its dependence on T alone, in which case it is too late to use the knowledge, or else calculate the marginal model, which seems to be as laborious as a full analysis. However, as we shall see, B may be able to profit from certain special structure in his model and prior to deduce that a statistic is marginally sufficient.

5. GENERALIZED SUFFICIENCY

It is not only, nor indeed primarily, the Bayesian who is motivated to simplify his problem of inference about Θ by discarding data. One obvious motivation for reducing the data to some statistic T is the possibility of eliminating nuisance parameters by satisfying the following definition:

Definition 5.1 (Basu, 1977). A statistic T is Θ-*oriented* if its sampling distribution is entirely determined by the value of Θ.

However, this property does not in itself justify one in discarding all the data but T, since one might be throwing away information relevant to inference about Θ. The Bayesian has, in marginal sufficiency, a coherent theory to tell him when he can reduce his data without essential loss. From the classical point of view, a variety of *ad hoc*, more or less intuitively reasonable ideas has been put forward, intended to identify properties of sampling distributions which serve to justify such reduction of the data.

A good account of these ideas is given by Barndorff-Nielsen (1978, Chapter 4). (See also Basu, 1977, 1978; Dawid, 1975) We shall concentrate on just two approaches, specializing Barndorff-Nielsen's definitions slightly.

5.1 *G-sufficiency*

This concept was introduced by Barnard (1963). The essence is as follows. Let the model be given by the family $P = \{P_\lambda\}$ of distributions for data X, and suppose these distributions are equivariant under the action of exact homomorphic transformation groups, G, acting on X, and $\bar{G}$ acting on Λ. That is to say, if $X \sim P_\lambda$, and $g \in G$, then $g \circ X \sim P_{\bar{g}\circ\lambda}$ (for further background see, for example, Dawid, Stone and Zidek, 1973).

Suppose the parameter of interest Θ is invariant under $\bar{G}$, so that $\Theta(\lambda) \equiv \Theta(\bar{g}\circ\lambda)$, and let T be the maximal invariant function of X under G. Then Barnard proposed that, in the absence of prior information, T should be regarded as containing all the available information about Θ. Such a statistic T is termed *G-sufficient* for Θ. It can be shown (see e.g. Lehmann, 1959, p.220) that, if Θ is a *maximal* invariant function of Λ under $\bar{G}$, then a G-sufficent statistic T will be Θ-oriented.

Example 5.1. Let $\mathbf{X} = (X^i : i = 1, \ldots, n)$ be a random sample from $N(\mu,\sigma^2)$. Take G as the additive group of real numbers, a typical element a taking $\mathbf{X}$ into $\mathbf{X} + a\mathbf{1}$; then we way may take $\bar{G} = G$, with $a \circ (\mu, \sigma^2) = (\mu + a, \sigma^2)$. A maximal invariant statistic is $\mathbf{X} - \bar{X}\mathbf{1} = (X^i - \bar{X} : i = 1, \ldots, n)$ (where $n\bar{X} = \Sigma_{i=1}^n X^i$), which is thus G-sufficient for the invariant parameter σ^2.

Further reduction is possible using ordinary sufficiency, either in the full or in the reduced model. Either way, this yields the statistic $\Sigma(X^i\text{-}\bar{X})^2$ in the above example, as containing all the available information about σ^2 in the absence of prior knowledge (about μ, in particular).

Example 5.2 Let $\mathbf{X}^i$ $(i=1, \ldots, n)$ be a random sample from the bivariate normal distribution with entirely unknown mean-vector and dispersion matrix. Let G consist of the group of location-scale transformations acting on each component separately (but identically for all i). After reduction by sufficiency, this yields the sample correlation coefficient as G-sufficient for its population counterpart.

Example 5.3. Sample Survey. Consider a sampling frame of labelled units, denoted by $i=1,2, \ldots ,m$. With unit i is associated an unknown quantity Y_i, and we take as our parameter Λ the ordered set $(Y_1, Y_2, \ldots , Y_m)$. The sampling scheme is determined by a known probability distribution P yielding S, a random subset of $\{1,2, \ldots ,m\}$, and the data consist of $X = \{(i, Y_i): i \epsilon S\}$.

Let G and $\bar{G}$ each be isomorphic to the group of permutations of $(1,2, \ldots , m)$, acting on data $x = \{(i,y_i): i \epsilon S\}$ as $g \circ x = \{(g^{-1}i, Y_i): i \epsilon S\}$, and on parameter $\lambda = (y_1, \ldots ; y_m)$ as $\bar{g} \circ \lambda = (y_{g1}, \ldots ; y_{gm})$. The sampling distributions are equivariant under G and $\bar{G}$ if and only if, under P, all subsets of the same size are equally probable; that is to say, for *simple random sampling* with a possibly random sample size. We have maximal invariants: $T=$the order statistic of $(Y_i: i \epsilon S)$, and $\Theta =$ the order statistic of $(Y_1, \ldots, Y_m)$, and thus, under simple random sampling, T is G-sufficient for Θ.

An ancillary statistic based on T is N, the size of sample taken, and the conditional distribution of T given (N,Θ) is a multivariate hypergeometric distribution.

Example 5.4. (Schou, 1978). Let $\mathbf{X}^i (i=1, \ldots, n)$ be a random sample of unit vectors in $\mathbf{R}^2$ drawn from the Fisher-von Mises distribution on the circle. The parameter Λ can take any value in $\mathbf{R}^2$, and the model densities have the form $c(\|\lambda\|) \exp (\lambda' \mathbf{x}) (\|\mathbf{x}\| = 1)$. A sufficient statistic is $\mathbf{S}_n = \Sigma_{i=1}^n \mathbf{X}^i$.

A typical element of G rotates each $\mathbf{X}^i$ about $\mathbf{0}$ through the same angle α, and has the same effect on both $\mathbf{S}_n$ and Λ. After a sufficiency reduction to $\mathbf{S}_n$, the maximal invariant is $\|\mathbf{S}_n\|$, which is thus G-sufficient for the maximal invariant parameter $\|\Lambda\|$.

5.2. *S-sufficiency*

Let T be a statistic. The experiment which yields observation of X, with

model densities $f(x|\lambda)$, can be regarded as made up of two components:

(i) the *reduced experiment*, yielding observation of T, with model densities $f(t|\lambda)$ derived by marginalization from $f(x|\lambda)$; and

(ii) the *conditional* experiment, after observing $T=t$, yielding observation of X but with model densities $f(x|t,\lambda)$ derived from $f(x|\lambda)$ by conditioning on T.

We suppose T is Θ-oriented, and try to express the fact the conditional experiment, discarded on reduction to T, contains no useful information about Θ. One such expression is the requirement that the conditional experiment is determined entirely by nuisance parameters. Because of the arbitrary nature of nuisance parameters, this may be interpreted *either* in terms of some nominated choice of nuisance parameter, *or* as a requirement that there exist *some* choice of nuisance parameter yielding this property. For non-triviality in this latter case we must impose some restrictions (otherwise Λ itself might be regarded as a nuisance parameter!) and this motivates the insistence that Θ and the nuisance parameter Φ should be *variation-independent*: that is to say, as Λ varies over its range of possible values, Θ and Φ range over a product-space. Thus the property of S-sufficiency may be expressed as:

$$f(x|\lambda) = f(t|\theta) f(x|t,\phi) \tag{5.1}$$

where Θ and Φ are variation-independent.
[Note that this property does *not*, in general, hold for G-sufficiency]

Example 5.5 Let X_1, X_2 have independent Poisson distributions with respective means Λ_1, Λ_2 known only to be positive. We are interested in $\Theta = \Lambda_1 + \Lambda_2$. Then $T = X_1 + X_2$ is Θ-oriented, and is in fact S-sufficient for Θ; for the conditional distribution of X given $T = t$ is Binomial $B(t;\Phi)$, where $\Phi = \Lambda_1/(\Lambda_1 + \Lambda_2)$ is variation-independent of Θ.

A trivial case of S-sufficiency arises when $X = (T,S)$, $\Lambda = (\Theta,\Phi)$ $(\Theta,\Phi$ variation-independent), and

$$f(t,s|\theta,\phi) = f(t|\theta) f(s|\phi). \tag{5.2}$$

Then the experiments producing T and S may be considered as entirely unrelated to each other.

Example 5.6. Components of variance. The data are $(X_{ij}: i=1, \ldots, I; j=1, \ldots, J)$, generated as

$$X_{ij} = \mu + \tau Y_i + \sigma Z_{ij}, \tag{5.3}$$

where the Y's and Z's are independent standard normal variables, and (μ, τ^2, σ^2) the value of the parameter. A minimal sufficient statistic is (S_1,S_2,S_3), where

$$S_1 = X_{..},\ S_2 = J\,\Sigma^I_{i=1}\ (X_{i.}-X_{..})^2,\ S_3 = \Sigma^I_{i=1}\ \ \Sigma^J_{j=1}\,(X_{ij}-X_{i.})^2,$$

and where the dot operator averages over the replaced suffix.

In the sampling distribution, S_1, S_2 and S_3 are mutually independent, with $S_1 \sim N(\mu, \sigma_0^2/IJ)$, $S_2 \sim \sigma_0^2\ \chi^2_{I-1}$, and $S_3 \sim \sigma^2\,\chi^2_{I(J-1)}$, where $\sigma_0^2 = \sigma^2 + J\tau^2$. Thus taking $\Theta = \sigma^2$, $\Phi = (\mu, \sigma_0^2)$, $T = S_3$, $S = (S_1, S_2)$, we have the factorization (5.2). However, Θ and Φ will not normally be variation-independent , since (with $\tau^2 \geq 0$) we must have $\sigma_0^2 \geq \sigma^2$, and it therefore seems that information in S may be relevant to σ^2. There are two ways in which we can get variation-independence: (1) restrict the parameter-space, for example requiring $\sigma^2 \leq a$ and $\sigma_0^2 \geq b\ (\geq a)$, μ being unrestricted; or (2) extend the parameter space to allow $\tau^2 < 0$. (This condition makes sense if interpreted in terms of the covariance structure of the (X_{ij}), rather than the synthetic representation (5.3): Dawid, 1977; we can then allow any combination of μ, $\sigma^2 > 0$, $\sigma_0^2 > 0$).

The former approach appears to distort the real problem to fit the Procrustean bed of theory, and in any case the appropriate implied parameter-space for (μ,τ^2, σ^2) will depend on the value of J. The latter approach may or may not be regarded as appropriate, leading as it does to the possibility of negative correlations between the $(X_{i.})$, and again involving the value of J.

The above problem is the subject of Stone and Springer (1965).

6. EXAMPLES OF MARGINAL SUFFICIENCY

Marginal sufficiency may or may not go hand in hand with its various classical counterparts, as the following examples illustrate.

Example 6.1. Full sufficiency. If T is sufficient for the full parameter Λ, then T is marginally sufficient for Θ for *any* prior distribution on Λ. Moreover, usually the converse will hold (Hájek, 1965; Martin, Petit et Littaye, 1973).

Example 6.2. G-sufficiency. In the model of 5.1, consider the family F of prior distributions for Λ which are *invariant* under $\bar{G}$; thus if $\Pi \in \mathbf{F}$, $g \in \bar{G}$, then $\Lambda \sim \Pi \Rightarrow \bar{g} \circ \Lambda \sim \Pi$. By general results on invariance (see e.g. Dawid, 1979a, Section 8), Θ is a "sufficient statistic" in F, so that F corresponds to a bevy of Bayesians, and thus leads to an agreed marginal model for X given Θ. It may now be seen that any of these distributions for X given Θ is invariant under the action of G on X, whence T is sufficient for this marginal model, and hence marginally sufficient. Thus all Bayesians in the bevy would agree to work with the marginal model for the reduced data T, and since T is, in any case, Θ-oriented, this is equivalent to using the sampling distributions of the G-sufficient statistic T.

In the context of Example 5.4, suppose that the prior distribution for Λ is rotationally symmetric about $\mathbf{0}$ (as a particular case, Λ_1 and Λ_2, might have independent standard normal distributions); then the posterior distribution of $\Theta = \|\Lambda\|$ will be a function of $T = \|\mathbf{S}_n\|$ alone, and could be derived by combining the marginal prior of Θ with the (Θ-oriented) reduced experiment for T.

Likewise, in Example 5.3 with simple random sampling, if in the prior distribution the variables $(Y_1,\ldots,Y_m)$ are *exchangeable* (which means, simply, invariance under the group $\bar{G}$ of permutations), then the order statistic T of the data will be marginally sufficient for the order statistic Θ of the parameter, and coherent inference could be based on its multivariate hypergeometric sampling distribution (for given sample size).

The general theory developed above is of somewhat limited applicability. A proper $\bar{G}$-invariant distribution exists only when $\bar{G}$ is compact as a topological group. Usually this condition does *not* hold; it fails, for instance, in Examples 5.1 and 5.2. Then $\bar{G}$-invariant measures exist, but are improper distributions. Difficulties can now arise. For example, it is possible for the posterior distribution of Θ to depend on the data through T alone, but not to be derivable from the reduced experiment based on T. This is the *marginalization paradox* of Dawid, Stone and Zidek (1973). Such problems do not arise for proper priors.

A difficult technical problem is to discover whether a G-sufficient statistic can be marginally sufficient for a non-invariant prior distribution, and, in particular, for a proper prior in the case of a non-compact group. Case studies suggest that this will not normally be possible (Jaynes, 1980). If so, then reduction to a G-sufficient statistic, when the group is not compact, will be intrinsically incoherent, in the sense that the only prior distributions which allow such reduction are improper, and possibly paradoxical.

Example 6.3. S-sufficiency. Suppose (5.1) holds, and the prior distribution for Λ is such that Θ and Φ are independent. (Thus, so long as the parameter-space is redefined, if necessary, as the support of the prior distribution, Θ and Φ must be variation-independent). Then $\pi(\theta,\phi) = \pi(\theta)\ \pi(\phi)$, whence

$$\pi(\theta,\phi\,|\,x) \propto \pi(\theta) f(t\,|\,\theta)\ \ \pi(\phi) f(x\,|\,t,\phi). \tag{6.1}$$

It follows that T is marginally sufficient for Θ, and the reduced experiment gives the marginal model.

In Example 5.5, suppose that we take a conjugate prior distribution: $\Lambda_i \sim \Gamma(a_i, b)$ independently. As is well known, this implies that $\Phi = \Lambda_1/(\Lambda_1 + \Lambda_2) \sim \beta(a_1, a_2)$, *independently* of $\Theta = \Lambda_1 + \Lambda_2 \sim \Gamma(a_1+a_2,b)$. It follows that $T = X_1 + X_2$ is marginally sufficient, so that inference for Θ follows on combining the reduced data T, having distribution $P(\Theta)$, with the marginal prior: $\Theta \sim \Gamma(a_1+a_2, b)$.

The above simplification is an important (but little-known) general property of conjugate inference for exponential families (Barndorff-Nielsen, 1978: Corollary 9.3). Under weak conditions, whenever a S-sufficient statistic T exists, yielding a factorization (5.1), then Θ and the nuisance parameter Φ will turn out to be independent, for any conjugate prior (where the term "conjugate"is suitably defined). Thus conjugate Bayes inference about such a parameter Θ can always proceed in the reduced experiment.

For Example 5.6, S_3 will be marginally sufficient for σ^2 (and (S_1, S_2) for (μ, σ_0^2)) if σ^2 and (μ,σ_0^2) are *a priori* independent. Again, interpreted in terms of (μ, τ^2, σ^2), this requirement cannot hold for more than one value of J, and so appears quite artificial.

Example 6.4. Complex sampling (Sugden, 1978). Suppose a sample survey is conducted as in Example 5.3, but with a complex sampling scheme which is not equivalent to simple random sampling. Consider again the family of exchangeable prior distributions, which constitute a bevy for inference about the order-statistic Θ of Λ, and hence yield an agreed marginal model for X given Θ. Once again, the order-statistic T of X is marginally sufficient for Θ; this follows because the posterior distribution does not depend on the sampling scheme, and since, for the particular case of simple random sampling, the posterior for Θ with an exchangeable prior depends on T alone, this must hold for any sampling scheme. Consequently, the bevy can confine itself to the reduced experiment for T.

Now in general T will not be Θ-oriented, and it would therefore seem that reduction of the data to T does not afford much simplification. However, it may be seen that simplicity returns if we work with the marginal model for T given Θ, as follows. Firstly, since sample-size N (a function of T) is ancillary in the full model, it is ancillary in the marginal model; and now a symmetry argument shows that, conditional on N, the marginal model for T will be multivariate hypergeometric, exactly as for simple random sampling.

Example 6.5. L-independence. (Barndorff-Nielsen, 1978, Example 3.8). Consider a birth and death process, with birth and death intensities Λ and M, observed continuously from time 0 to time T, in which there are initially ℓ individuals. Let B, D and Z denote respectively the number of births, the number of deaths, and the total time lived by all individuals. Then (B, D, Z) is sufficient for (Λ, M), and the likelihood based on data (b, d, z) is proportional to

$$\lambda^b \mu^d e^{-(\lambda+\mu)z}. \tag{6.1}$$

Since this factorizes as a function of λ and μ, we call Λ and M *L-independent*, although (6.1) can *not* be produced by S-sufficiency, and is not of the form (5.1).

Suppose that Λ and M are *a priori* independent. Then $\pi(\lambda \mid \text{data}) \propto \pi(\lambda). \lambda^b e^{-\lambda z}$, a very straightforward calculation. For inference about Λ, all the Bayesian has to do is to store the relevant factor of his likelihood and combine it with his prior.

Here $T = (B, Z)$ is marginally sufficient for Λ, but it would not be quite so straightforward to make inference about Λ from the reduced experiment, since (B, Z) is *not* Λ-oriented and has a complicated distribution. In this case, it does not help to derive the marginal model for (B, Z) given Λ, which is also complicated and depends on the distribution assigned to M.

The lesson here is that, even when a marginally sufficient statistic exists, it may not be most profitable to the Bayesian to work with its sampling distributions (in either the full or the marginal model); other uses may be more appropriate. The same moral is pointed by the next example.

Example 6.6. Optional stopping. Consider again the Fisher-von Mises distribution of Example 5.4, but with sequential observation of $\mathbf{X}^1, \mathbf{X}^2, \ldots$, stopping according to the following rule: if $\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^r$ have been observed with values $\mathbf{x}^1, \ldots, \mathbf{x}^r$, then observations terminates if the first component x_1^r of $\mathbf{x}^r$ is negative; otherwise $\mathbf{X}^{r+1}$ is observed. This rule leads, with probability one for all Λ, to termination of observation at some random finite stage N.

The data may be expressed as $(n,\mathbf{x}^1, \ldots, \mathbf{x}^n)$, the observed values of $(N, \mathbf{X}^1, \ldots, \mathbf{X}^N)$. By a standard result on optional stopping, the posterior distribution for Λ will be identical with that based on observing values $(\mathbf{x}^1, \ldots, \mathbf{x}^n)$ for $(\mathbf{X}^1, \ldots, \mathbf{X}^n)$ in the non-sequential set-up of Example 5.4, for the appropriate value of n.

In particular, consider the bevy of prior distributions for Λ which are rotationally symmetric about $\mathbf{0}$. Then, by the results of Example 6.2, the posterior distribution of $\Theta = \| \Lambda \|$ will depend only in the value of $(N, \|\mathbf{S}_N\|)$ (the value of N, taken for granted as known earlier, must now be specified). As in the last two examples, the marginally sufficient statistic $(N, \|\mathbf{S}_N\|)$ will not in general be Θ-oriented (the non-invariant stopping rule destroys that property), and so the bevy might wish to focus attention on the marginal model for $(N, \|\mathbf{S}_N\|)$. It might be conjectured, in analogy with Example 6.4, that N is ancillary in this marginal model, and that conditioning on it produces the same distribution for $\|\mathbf{S}_N\|$ as in Example 5.4. However, N is not ancillary. For example, it may easily be seen that, for $\Theta = \mathbf{0}$ (which gives an uniform distribution on the circle), the distribution for N given Θ is geometric with probability parameter 1/2; while for Θ very large, corresponding to the $(\mathbf{X}^i)$ being tightly concentrated about the same random unit vector $\mathbf{e} = \Lambda/\Theta$,$N$ will tend to be either 1 (if $e_1 < 0$) or otherwise very large (if $e_1 > 0$); each extreme holding with probability about 1/2.

Consequently, conditioning the marginal model on N is inappropriate, and we do not recover the same reduced marginal model as for Example 5.4. It seems that our bevy cannot shortcut the complicated task of calculating the reduced marginal model.

However, this is, in reality, quite unnecessary. We know that posterior distributions will be identical with those for Example 5.4, which are easily found, so that use of the marginal model may be completely by-passed. Alternatively, we might say that it is in order to use an entirely fictitious model, in which $N = n$ is regarded as fixed and $(\mathbf{X}^1, \ldots, \mathbf{X}^n)$ drawn as a random sample of size n. Once again, we have a simple marginally sufficient statistic leading to a simple Bayesian inference, but it is not all helpful to work with sampling distributions.

7. D-SUFFICIENCY AND D-ANCILLARITY

The examples of Section 6 demonstrate that, even when a simple marginally sufficient statistic T exists, leading to a simple marginal posterior distribution for Θ, it may well not be fruitful for the Bayesian to concern himself with the sampling distribution of T. In particular, whether or not T is Θ-oriented will depend on irrelevant properties of the sample-space (compare Examples 6.4 and 6.6 with Example 6.2). Consequently, our next definition may be of little interest to the whole-hearted Bayesian.

Consider a model for data X, with parameter Λ, and a Bayesian B with prior distribution Π for Λ

Definition 7.1. A statistic T is *D-sufficient* for Θ (for B, or Π) if T is (i) marginally sufficient for Θ, for B, and (ii) Θ-oriented.

This definition is important for purposes of comparing Bayesian and classical concepts. In particular, we shall be examining the classical prescriptions for reduction to T, which do depend on the sample space and do, usually, have T Θ-oriented, to discover when they can be given a Bayesian justification.

From the classical viewpoint, there is another common way of eliminating nuisance parameters, namely by *conditioning*. This involves replacing the original experiment for X by the conditional experiment for X, given a statistic T. In parallel with reduction, this is motivated by the possibility of achieving the following simplification.

Definition 7.2. A statistic T is Θ-*inducing* if, for any t, the conditional experiment for X given $T = t$ is determined entirely by the value of Θ.

Using only the conditional experiment involves discarding the reduced experiment for T, and we therefore require criteria which allow us to do so without losing "useful information". These are entirely analogous to the criteria involved in discarding a conditional experiment, as already considered, and the two problems are in effect two faces of the same coin, labelled "non-formation" by Barndorff-Nielsen (1976, 1978).

We shall specifically consider the following criterion.

Definition 7.3. A (Θ-inducing) statistic T is *S-ancillary* for Θ if there exists a nuisance parameter Φ, variation-independent of Θ, which determines the reduced experiment for T (which is to say that T is Φ-oriented).

A *S*-ancillary statistic T gives rise to the factorization

$$f(x|\lambda) = f(x|t,\theta)f(t|\phi). \tag{7.1}$$

Comparing this with (5.1), we see that T is S-ancillary for Θ if and only if T is S-sufficient for Φ. Thus, in Example 5.5, $T = X_1 + X_2$ is S-ancillary for $\Phi = \Lambda_1/(\Lambda_1 + \Lambda_2)$, and this might justify basing inference about Φ on the conditional (binomial) model for X given T.

For the Bayesian, a generalized ancillarity criterion, which would allow him to work with a conditional model,rather than the full model , seems even less worthy of attention than generalized sufficiency, since he is not normally concerned with sampling models anyway, and in this case does not even gain, in general, by being able to discard data. One again, the following definition is of most importance for purposes of comparison between Bayesian and classical ideas.

Definition 7.4. A statistic T is *D-ancillary* for Θ (for B, or Π) if it is (i) marginally ancillary for Θ, for B, and (ii) Θ-inducing.

(Recall that "T is marginally ancillary for Θ" means that T is an ancillary statistic in the marginal model, so that $\bar{f}(t|\theta)$ does not depend on θ).

If T is D-ancillary for Θ, then $f(x|\lambda) = f(x|t, \theta) f(t|\lambda)$, whence $\bar{f}(x|\theta) = \int f(x|\lambda)\, \pi(\lambda|\theta)d\lambda = f(x|t,\theta) \int f(t|\lambda)\, \pi(\lambda|\theta)d\lambda = f(x|t,\theta)\, \bar{f}(t|\theta) \propto f(x|t,\theta)$. It follows that the posterior distribution for Θ satisfies $\pi(\theta|x) \propto f(x|t,\theta)\, \pi(\theta)$, and so can be found by combining the prior marginal distribution for Θ with the conditional model given T. Conversely, when T is Θ-inducing, marginal ancillarity of T is necessary for this property to hold. Thus D-ancillarity may be regarded as a Bayesian justification for working with the conditional model. Note once again that the definition involves only the *conditional* prior distribution for Λ given Θ, and so is relevant for the whole bevy of Bayesians sharing this conditional distribution, the marginal prior distribution for Θ being arbitrary.

Suppose T is S-ancillary for Θ, so that (7.1) holds. Trivially, if Θ and Φ are *a priori* independent, then $\pi(\phi|x) \propto \pi(\phi) f(x|t,\phi)$, so that T is marginally ancillary. The use of the conditional model is thereby justified if the prior independence holds. Again, it will normally hold for conjugate inference in exponential families.

The following example (from Dawid and Dickey, 1977) shows that prior independence is *not* necessary for a S-ancillary statistic to be D-ancillary.

Example 7.1. Suppose $f(x|\lambda) = f(x|t,\theta) f(t|\phi)$, where (Θ,Φ) takes values in $[-1,1] \times [-1,\frac{1}{2}]$. We need not specify $f(x|t,\theta)$, but suppose $f(t|\phi) = 2t^{-3} (1 + \phi\, t) / g(\phi)$ for $t \geq 1, -1 \leq \phi t \leq ½$; 0 otherwise. The normalizing constant is

$$\begin{aligned} g(\phi) &= (1+\phi)^2 \quad (-1 \leq \phi \leq 0) \\ &\ (1+2\phi-8\phi^2)\ (0 \leq \varphi \leq ½). \end{aligned}$$

In the prior, Θ and Φ are *not* independent, and in fact

$$\pi(\phi|\theta) = (4/3)\,(1-\theta\phi)\, g(\phi)\ (-1 \leq \phi \leq ½).$$

(This does define a density, for any $\theta \in [-1,1]$.)
We find $\bar{f}(t|\theta) = \int f(t|\phi)\, \pi(\phi|\theta)\, d\phi = 3t^{-4}\ (t \geq 1)$
$\qquad\qquad\qquad\qquad 0 \quad$ (otherwise)

so that T is both S- and D-ancillary for Θ.

(A similar example may be constructed to show that statistic T may be both S- and D-sufficient for Θ, although Θ and Φ are not *a priori* independent).

In the next Section we examine in more detail the connexions between S- and D-sufficiency and ancillarity.

8. S- AND D-NONFORMATION

The material of this Section draws heavily on Dawid and Dickey (1977).

The concepts of sufficiency and ancillarity being considered may usefully be expressed in the general framework of *conditional independence* (Dawid, 1979a), and our theorems below are applications of general properties of conditional independence to our specific problems. For further technical background and rigorous proofs, see Dawid (1980).

8.1. Ancillarity

Suppose we have S-ancillarity: $f(x|\lambda) = f(x|t,\theta) f(t|\phi)$. Suppose further that T *strongly identifies* Φ, as defined in Dawid (1980): that is to say, if we consider the marginal distribution of T induced by assigning a prior distribution to Φ, two different priors will induce distinct marginal distributions. This property is commonly known as "identification of mixtures" (Teicher, 1960, 1961, 1967; Barndorff-Nielsen, 1965; Chandra 1977). Clearly, strong identification implies ordinary identification.

Theorem 8.1. Under the above conditions, T is D-ancillary for $\Theta \Leftrightarrow \Theta$ and Φ are *a priori* independent.

Proof. We have already shown "$\Leftarrow$". For "$\Rightarrow$", we note that $\bar{f}\,(t|\theta) = \int f(t|\phi)\ \pi(\phi|\theta) d\phi$, and marginal ancillarity gives that $\bar{f}(t|\theta_1) = \bar{f}(t|\theta_2)$ for any θ_1, θ_2. Since $\bar{f}(t|\theta)$ is a mixture of $f(t|\phi)$ with mixing measure $\pi(\phi|\theta)$, strong identification implies that $\pi(\phi|\theta_1) = \pi(\phi|\theta_2)$, so that we have independence.

We can summarize this result as saying that, with the strong identification property, use of S-ancillarity to allow inference from the conditional model is coherent (i.e. has a Bayesian justification) if and only if Θ and Φ are *a priori* independent; more informally, it is necessary and sufficient that Θ and Φ each carry no information about the other.

The next result gives conditions on the prior distribution, not involving the model, under which S- and D-ancillarity can *never* co-exist.

Theorem 8.2. Suppose that the prior conditional distributions of Φ given Θ, considered as a parametric family, are *boundedly complete*; that is, if $h(\Phi)$ is bounded with $E[h(\Phi)|\Theta] = 0$ a.s., then $h(\Phi) = 0$ a.s. If (7.1) holds, then T is *not* D-ancillary for Θ.

Proof. Suppose the contrary, and let $k(T)$ be a bounded function. Since T is Φ-oriented, $E[k(T)|\Theta,\Phi] = E[k(T)|\Phi] = h(\Phi)$ say. Then $E[h(\Phi)|\Theta] = = E[k(T)|\Theta] = E[k(T)]$ a.s. since T is marginally ancillary. So by bounded completeness $h(\Phi) = E[k(T)]$ a.s., i.e. $E[k(T)|\Theta,\Phi]$ = constant a.s. As this holds for any k, T must be independent of (Θ,Φ), so that T is in fact ancillary, and so cannot be Θ-inducing (barring the trivial case that X is Θ-oriented).

8.2. *Sufficiency*

Suppose we have S-sufficiency: $f(x|\lambda) = f(t|\theta)f(x|t,\phi)$. We look for a result analogous to Theorem 8.1.

Theorem 8.3. If, for each value of t, the distributions of X given $T = t$ strongly identify their parameter Φ, then T marginally sufficient for $\Theta \Rightarrow \Theta$ and Φ are independent in their distribution posterior to observing T.

The proof parallels that of Theorem 8.1.

Under the strong identification condition of Theorem 8.3, the distribution of Φ given (T,Θ) does not depend on Θ. Also, since T is Θ-oriented, T is independent of Φ given Θ, so that the distribution of Φ given (T,Θ) does not depend on T. We appear to have shown that Φ is independent of (T,Θ), and thus that Θ and Φ must be independent *a priori*. However, this reasoning is fallacious without further conditions (Dawid, 1979b).

Example 8.1. The parameter is (Θ,Φ) with $\Phi > 0$, $\Theta \neq 0$. The data are $(S,T) = (Y/\Phi, Z/\Theta)$, where Y and Z have independent standard exponential distributions, with density $f(y) = e^{-y}(y>0)$. We thus have "unrelated problems". Given T, the data X reduce to S, with distribution unchanged, and S strongly identifies Φ, by the uniqueness property of the Laplace transform.

Suppose the prior distribution has

$$\pi(\phi|\theta) = \begin{array}{l} e^{-\phi}\ (\phi>0) \text{ when } \theta > 0 \\ 2e^{-2\phi}(\phi>0) \text{ when } \theta < 0. \end{array}$$

Then T is D-sufficient for Θ; indeed, we may take

$$\bar{f}(s|t,\theta) = \begin{array}{l} (1+s)^{-2}\ (s>0) \text{ when } t > 0 \\ 2(2+s)^{-2}\ (s>0) \text{ when } t < 0 \end{array}$$

independently of θ. However, Θ and Φ are *not* independent in the prior distribution.

The further condition needed to ensure the validity of our informal argument above is the non-existence of a set A for which $P(T \epsilon A|\theta)$ is always 0 or 1, both values being taken as θ varies. Such a set is called a *splitting set* for T given Θ (Koehn and Thomas, 1975). In Example 8.1, the positive half-line is such a splitting set.

We thus have the following result.

Theorem 8.4. Suppose T is S-sufficient for Θ and that there does not exist a splitting set for T given Θ. Suppose further that, for each value of t, the distributions of X given $T = t$ strongly identify the nuisance parameter Φ. Then T is D-sufficient for Θ if and only if Θ and Φ are *a priori* independent.

Thus, under appropriate conditions on the model, reduction by S-sufficiency is "coherent" if and only if Θ and Φ are *a priori* independent. (Note that this result, in common with Theorem 8.1, does not use the property that Θ and Φ be variation-independent).

Example 8.2. In the components of variance problem of Example 5.6, take $\Theta = (\mu,\sigma_o^2)$, $\Phi = \sigma^2$, $T = (S_1, S_2)$, $S = S_3$. Then the conditions of Theorem 8.4 hold, so that inference for (μ,σ_o^2) based on (S_1,S_2) alone is coherent if and only if (μ,σ_o^2) is *a priori* independent of σ^2: a condition which, as indicated earlier, is unrealistic. The same condition is necessary and sufficient for coherent inference about σ^2 based on S_3 alone.

In this example, interest may well centre on μ alone, so that reduction to (S_1,S_2) would not eliminate all nuisance parameters. It seems likely that (S_1,S_2) will be marginally sufficient for μ (although not, of course, μ-oriented) only under the above prior independence; however, I do not have a proof of this.

Stone and Springer (1965, Rider) prove a theorem very similar to Theorem 8.4 and apply it to the variance-components model. However, they omit the splitting-set condition.

Example 8.3. We show that the strong identification condition of Theorem 8.4 may not always be required for the result to hold. Consider again Example 5.5, and suppose T is D-sufficient for Θ. We have

$$\bar{f}(x_1|t,\theta) = \binom{t}{x_1} \int_0^1 \phi^{x_1} (1-\phi)^{t-x_1}\, \pi(\phi|t,\theta)d\phi$$

and $\pi(\phi|t,\theta)$ may be replaced by $\pi(\phi|\theta)$, since T is Θ-oriented, so that T and Φ are independent given Θ. Thus $\bar{f}(x_1|t,\theta)$ will be determined by the first t moments of $\pi(\phi|\theta)$, and, so long as these are the same for every value of θ, $\bar{f}(x_1|t,\theta)$ will not involve θ. Here the distributions of X given $T = t$ do *not* strongly identify Φ, for any t. However, the marginal sufficiency requirement that $f(x_1|t,\theta)$ should not involve θ *for all* t ensures that *all* moments of $\pi(\phi|\theta)$ are constant, whence $\pi(\phi|\theta)$ is itself constant, so that we must have Θ and Φ independent.

REFERENCES

BARNARD, G.A. (1963). Some logical aspects of the fiducial argument. *J. Roy. Statist. Soc., B.* **25**, 111-114.

BARNDORFF-NIELSEN, O. (1965). Identifiability of mixtures of exponential families. *J. Math. Anal. Appl.*, **12**, 115-21.

— (1976). Nonformation. *Biometrika* **63**, 567-571.

— (1978). *Information and Exponential Families in Statistical Theory*. Wiley: Chichester - New York - Brisbane.

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Ass.* **72**, 355-366.

— (1978). On partial sufficiency: a review. *J. Stat. Plann. Inference*, **2**, 1-13.

CHANDRA, S. (1977). On the mixtures of probability distributions. *Scand. J. Statist.* **4**, 105-112.

DAWID, A.P. (1975). On the concepts of sufficiency and ancillarity in the presence of nuisance parameters. *J. Roy. Statist. Soc. B.* **37**, 248-258.

— (1977). Invariant distributions and analysis of variance models. *Biometrika* **64**, 291-7.

— (1979a). Conditional independence in statistical theory (with Discussion). *J. Roy. Statist. Soc. B* **41**, 1-31.

— (1979b). Some misleading arguments involving conditional independence. *J. Roy. Statist. Soc. B* **41**, 249-252.

— (1980). Conditional independence for statistical operations. *Ann. Statist.* **8**, 598-617.

DAWID, A.P. & DICKEY, J.M. (1977). Problems with nuisance parameters-traditional and Bayesian concepts. *Tech. Report.*, University College London.

DAWID, A.P., STONE, M. & ZIDEK, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference (with Discussion). *J. Roy. Statist. Soc. B*, **35**, 189-233.

HAJEK, J. (1965). On basic concepts of statistics. *Fifth Berkeley Symposium on Mathematical Statistic and Probability* 1, 139-162.

JAYNES, E.T. (1980). Marginalization and prior probabilities. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, (A. Zellner, ed.). Amsterdam: North Holland.

KOEHN, U. & THOMAS, D.L. (1975). On statistics independent of a sufficient statistic: Basu's lemma. *American Statistician* **29**, 40-42.

LEHMANN, E.L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.

LINDLEY, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*. Cambridge: University Press.

MARTIN, F., PETIT, J.L. & LITTAYE, M. (1973). Indépendance conditionelle dans le modèle statistique bayésien. *Ann. Inst. Henri Poincaré, B*, **9**, 19-40.

RAIFFA, H.A., & SCHLAIFER, R.S. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University.

SCHOU, G. (1978). Estimation of the concentration parameter in von Mises-Fisher distributions. *Biometrika* **65**, 369-377.

STONE, M. & SPRINGER, B.G.F. (1965). A paradox involving quasi prior distributions. *Biometrika* **52**, 623-627.

SUDGEN, R.A. (1978). *Exchangeability and the foundations of survey sampling*. Ph. D. Thesis, University of Southampton.

TEICHER, H. (1960). On the mixture of distributions. *Ann. Math. Statist.* **31**, 55-73.

— (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32**, 244-248.

— (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.* **38**, 1300-2.

DISCUSSION

G.A. BARNARD (*University of Waterloo, Canada*):

I welcome very much Professor Akaike's presence at this conference; not only because I have not before had the opportunity to discuss issues with one whose contributions to statistics have been so constructive and important, and I welcome this now; but most of all because I have been coming to think that the apparent division of statisticians into doctrinally opposing 'schools' of Bayesian and anti-Bayesians was doing great harm to our subject, and would do more if it was allowed to continue. The purity of doctrine of the organising Committee of this Conference is beyond question; also beyond question is the fact that Professor Akaike is not a subscriber to the pure doctrine. The fact that he was invited, and accepted, to speak here is therefore specially welcome.

To deal first with a minor point. I cannot go along with Professor Akaike's criticism of Savage's axioms, any more that I think Wolfowitz's criticisms were justified. Surely the young boy's difficulty arises from his regarding the difference of ranks by 1 as marginal. Unless he moves himself and his friends to a Muslim country he will have to decide eventually; and his eventual choice may always be supposed to arise from a perception that one, at least, of the 'marginal' differences is, in fact, more important than the others.

Wolfowitz's criticisms seemed to me misdirected, because if we have a *general* rule for choosing one from among any set of decision rules, we must be able to rank the set.

However, I would go along with a slight re-formation of Professor Akaike's argument. For the 'inconsistent triad' to which his argument leads arises also in the Arrow-Condorcet Theorem concerning the impossibility of collective 'democratic' choice. And I agree that statistics is nothing if it is not concerned with *objective* analyses of data, in some sense -- whatever we may care to say about the 'nature' of the probabilities with which we deal these probabilities must be *agreed* between several people. And it follows that Savage's argument does not serve to demonstrate the necessary existence of such *agreed* propabilities for any proposition we care to think of. We cannot therefore take for granted the existence of agreed prior probabilities for all the parameters involved in a model of an experiment. But,such agreed priors are necessary for the universal applicability of Bayes'Theorem.

Professor Akaike appears to accept what I call the Likelihood Model (LM) as typical of the logical structure of an experiment. This specifies the sample space $S = [x]$ of possible results, the parameter space $\Omega = [\theta]$ of possible parameter values, and the probability function $f(x,\theta)$ giving the probability of x when the parameter value is θ. Given the three elements $[S,\Omega,f]$ we can deduce the distribution of the likelihood function and from it derive, at least in some cases, a 'prior' distribution for θ in accordance with Professor Akaike's principles. If we then represent the inference by the posterior distribution of θ relative to this prior and the observed likelihood function we shall obtain an inference which has a clear frequency interpretation. I assume that Professor Akaike would accept that such an inference is appropriate only when we really have no observational basis for any statement about the parameter values other than that which is implicit in the design of the experiment.

We should not, I think, be over-ready to assume we are in this state of ignorance. For example, as Akaike shows, his rule can be made to yield the rules of ridge regression. But this, in spite of its recent vogue, by no means always gives an improvement on standard least squares. A group of statisticians in a large chemical company were persuaded to review a sample of experiments they had analysed in the past, in cases where the true values of the parameters had become essentially known. It turned out that ridge regression was, on the whole, worse than ordinary least squares for these cases. A reason for this could be found in the fact that ridge regression can be seen to be equivalent to an assumption that the parameters being estimated themselves follow a spherical normal distribution centred on the origin. Thus if all parameters in an experiment are roughly of the same order of magnitude, and their signs are randomly distributed, we can expect ridge regression to improve over ordinary least squares. But if, as with the experiments reviewed, one or two of the parameters were very large, while the remainder were quite small, ridge regression would not do so well. Thus background knowledge of the kinds of parameter value likely to be met with should be used in addition to such information about the prior as may be deduced from the experimental design. This last, in fact, can only be supposed to convey information about the prior in so far as it reflects the knowledge of the experimenter.

Thus I believe that when we use a Bayesian model for the analysis of an experiment we should regard the prior distribution in much the same way as we regard the form assumed for the probability function as something which may perhaps be important to our inference, so that we should be careful to check how far this is so; and as something which is capable of objective verification, at least in a long run and which, of course, we should so verify. This long run verifiability is, I think, the source of the objectivity, such as it is, which may be claimed for a Bayesian analysis.

In appealing to long run verifiability, of course, we need to specify the long run we consider relevant, the class of experiments to which we judge the current one belongs. A chemical engineer will find little difficulty in viewing the current chemical reaction rate constant which he is measuring as one of a set of such rates; and for the given equipment which he has available he will have had to set his temperature and other features of his design so that the reaction rate is neither too fast nor too slow to be measurable. He will then not go far wrong in using a prior distribution which is reasonably uniform over the range of measurable values. Similar considerations apply to an econometrician measuring elasticities, etc. But a physicist who is measuring the velocity of light, or some fundamental natural constant, would find it hard to regard his parameter as just one of a class of such, following a distribution concerning which he has any knowledge at all. This is why, it seems to me, Bayesian models are appropriate for experiments in chemical engineering, or in econometrics —provided, of course, the conclusions are understood as being subject to the correctness, to sufficient approximation, of the prior assumptions— but they are less appropriate for fundamental work in physics.

Professor Dawid has, as usual, presented us with a paper which stimulates us to further examination of foundations. But he omits, I think, to question a presupposition which ought to be questioned: Do nuisance parameters, as now commonly understood, exist? How often can we, or should we, ‘ignore’ parameter that enters into the specification of our experimental model?

When Hotelling first introduced the term, ‘nuisance parameter’ meant just what it

said: a parameter that one would have preferred to omit from one’s model. But now the term has come to have an unfortunate use among the adherents of what (at Barndorff-Nielsen’s suggestion) may be called the ‘prespecification school’ of mathematical statisticians. By this is meant the school which, given a model for an experimental situation, lays down *in advance of the data* the *kind* of conclusion that is to be reached. In relation to nuisance parameters, a typical requirement is that a ‘test’ of prespecified size should be provided that is ‘unbiased’ or ‘similar’. Such prespecified requirements can easily lead to absurdity. The 2x2 table:

	A	*not-A*	*Total*
Population I	*a*	*b*	*m*
Population II	*c*	*d*	*n*
Total	*r*	*s*	*N*

provides a simple example. If p_1,p_2 are the probabilities of A in populations I and II respectively, we often are interested in the crossratio parameter $\theta = p_1q_2/p_2q_1$, where $q_i = 1\text{-}p_i$, $i = 1,2$, and less interested in the ‘nuisance parameter’ $(p_1+p_2)/2$ which we may denoted ϕ. If we now prespecify (as is done, for example, in Lehmann‘s book) a test of $\theta = 1$ of size (say) 0.05, which is to be similar, or unbiased, against alternatives $\theta \neq 1$, we must reject the hypothesis tested with probability not less that 0.05 when $p_1 = 2\times 10^{-10}$ and $p_2 = 10^{-10}$. But when this is the case we will, with practical certainty, get the result $a=0$, $b=m$, $c=0$, $d=n$, and so we must reject, given this result, with probability 0.05. But to reject at all with such a result is clearly absurd.

One of the biggest advantages of the Bayesian approach over that of the prespecification school is that the Bayesian model gives a primary inference in the form of a posterior distribution for all the parameters involved in the experimental model. *If*, for example, the posterior distribution is very nearly normal, then it may be judged reasonable to express the conclusion in terms of an ‘estimate’ with a standard error; but whether this will be so, or not, may well depend on the data as well as on the model and the prior. And it may turn out that we can express the posterior in terms of two parameters θ and ϕ such that, a posteriori, these two variables are, to a sufficient approximation, independent. In such a case we can treat ϕ as a ‘nuisance parameter’ in relation to θ; but to do this in other cases can be dangerous. Certainly the mere fact that we would *like* to make an inference about θ without referring to ϕ by no means implies that we can. If we insist on doing so we are guilty of adopting the ‘prespecification’ approach.

Thus Dawid’s statement (p.5) that ‘From the point of view of the single Bayesian B, the marginal likelihood is as good as any ordinary likelihood. . .’ needs qualification. If θ and ϕ, given x, are far from independent, then further information which may well come to hand concerning ϕ will affect the conclusions we draw concerning θ; and the mere statement of the marginal distribution of θ will give no expression to this fact. By contrast, if we have an ‘ordinary’ likelihood for θ — that is, one from an experiment in which θ alone is involved — then no further information about another parameter ϕ alone will affect our conclusion, provided θ and ϕ are independent a priori. I am, of course, assuming here something which I regard as fundamental to natural science — the possibility of *knowing* that two distinct experiments are independent of each other.

I am led to wonder whether the term 'nuisance parameter' should not be left to the exclusive use of the prespecification school. I have long thought it unfortunate that Student's t statistic should have such beautiful properties that we are too often tempted to make inferences (posteriors or confidence distributions) which relate to location μ only, when in fact we almost always ought to make simultaneous inferences about location μ and scale σ together. Terms like 'primary' and 'secondary', to indicate a *ranking* of our interest, rather than a total lack of interest, would usually be more appropriate. George Box has introduced the term 'discrepancy parameter' to describe the kind of parameter in which our interest is minimal. The concept which I would like now to discuss is a little different, I think, and I shall use the term 'model adjustment parameter'. I hope I will not be drummed out of the Conference if I describe the idea in connection with a 'classical', not-necessarily-Bayesian problem:

We are given two samples, of size m, n respectively, from normal populations with means μ_1, μ_2, and standard deviations σ_1, σ_2, all unknown. We want to test whether $\mu_1 = \mu_2$ or not. In the text books we are told that if it is known that $\sigma_1 = \sigma_2$, then we can reduce the problem to a t-test; but if it is now known that $\sigma_1 \neq \sigma_2$ then, typically, we are told the problem is 'difficult'. We may, or may not, be referred to Fisher's tables, or to Welch or to Gurland. What never occurs, so far as I can tell, is an invitation to set $\lambda = \sigma_2/\sigma_1$ and then put

$$t(\lambda) = (\bar{y}\text{-}\bar{x}) \;/\; \surd[((1/m) + (\lambda^2/n))\;((m\text{-}1)s_x^2 + (n\text{-}1)s_y^2/\lambda^2)/(m+n\text{-}2)]$$

In many cases, if we plot $t(\lambda)$ over the plausible range of λ we shall find that it varies only trivially. In this case, we can make our inference about the difference between μ_1 and μ_2 knowing that it will be unaffected by the possible difference in the variances. If things turn out otherwise, then, but only then, we must either obtain more information about λ, or we must resort to Behrens-Fisher, or some such type of argument. I myself cannot recall a practical case where the inference was seriously affected by λ.

The parameter λ here plays the role of what I propose to call a 'model adjustement' *(MA)* parameter. This is a parameter which needs to be specified in order to define the distributions involved in our experiment, but which varies over a relatively narrow range. We have reason to suppose that our inference will, with high probability, turn out not to depend on that value of the *MA* parameter. Should we be disappointed in this hope, then we must either record the fact that our inference depends on the value taken for this parameter, or we must find out more precisely what value this parameter really takes. Finally we may use some special form of argument to derive an inference which *explicitly* depends on ignorance of the *MA* parameter value.

Typical of the arguments 'from ignorance' here referred to is that involved in the derivation of the Behrens-Fisher test, where we have a pivotal quantity $s_x^2\lambda^2/s_y^2$ for the parameter concerning which we wish to express our ignorance. We condition on the observed ratio s_x^2/s_y^2 and *conventionally* retain the distribution of the pivotal by a conceptual distribution of the parameter λ. Statements we then make concerning the other parameters must be interpreted as referring to the reference set thus conceptually generated. Whether or not such modes of reasoning come to be generally understood and accepted

will largely depend on whether the results they give appear 'reasonable' to the scientific community at large; in this respect the conventions as to the interpretation of 'ignorance' thus introduced may be compared with such conventions as the interpretation of the terms 'set', 'class', etc., in the foundations of mathematics. Most mathematicians seem to agree that the results derivable using the 'axiom of choice' correspond to their 'intuition' of the sort of structure that mathematics ought to be. It is known that the axiom of choice can be negated without thereby introducing a contradiction into set theory;but systems built on such negation seem in some sense 'pathological' to most mathematicians. To put the matter loosely, when we ask about the difference of means when both location parameters are unknown, we are asking a slightly silly question; and we must be content with a slightly silly answer. Certainly the answers thus arrived at, subject to marginalization paradoxes though they be, seem to me to have as much Bayesian justification as the answers obtained by simple marginalization from a posterior involving the secondary, *MA*, or nuisance parameter if we then forget that our conclusions concerning the parameter of primary interest are subject to modification should further data become available concerning the parameter we have integrated out.

Thus I agree with Zellner's comment concerning 'marginalization paradoxes'. We must remember that statistics is intended for application to the advancement of science. Fanatical insistence on freedom from 'incoherence' can lead to such complicatedly interrelated analyses of data as to go well beyond the capacity of our understanding. Judicious simplifications are an essential component of scientific advance.

P.R. FREEMAN (*Leicester University*):

When I first read Professor Akaike's paper I thought "If he goes to Spain and reads that, he'll be a brave man indeed". Well, he has - and he is. How can I react? I could fill all the discussion time with an uptight, strict Bayesian reply but this would be too negative. I must first, though, say that I can see no force in the counterexample to Savage's axiom of choice and that only very rarely (as in weather forcasting) am I at all interested in the expected performance of a Bayesian procedure. I can't therefore see any sense in the argument of section 2.3 and would happily condemn to the statistical mental asylum anyone who needed to know whether sampling was going to be direct or inverse before stating his prior for θ. Similarly, after many close readings of section 3 I am still not clear exactly what "objectivity" is claimed for the likelihood function and prefer to stick to the viewpoint of that great statistician Shakespeare (1598) who said

But (by your leave) it never yet did hurt,
To lay down likelihoods and forms of hope.

Likelihoods are, to me, just as much "forms of hope" as any other ingredients in the inference mixture.

To be more positive, let me turn to matters on which we can agree wholeheartedly. I take Professor Akaike's point to be that there are more things in real analysis than are dreamed of in any of our statistical philosophies. There must always be a rather messy interplay between the data and the choice of model, of parameters and of priors on those parameters if our analyses are to be of any value at all. This paper presents some very ingenious ways in which this can happen and they all show great promise in the applications we see. But we can all do quite well (well, nearly all, my own paper being one exception) when we generate an artificial set of data with known parameter values, know we are using the correct model and furthermore reuse the data to choose the best prior for us. Figs. 3 and 4 are the only ones relating to real data, so I should like to see several more real examples before judging the results.

To me, the two fundamental questions raised by this paper are:

i) Do these ideas give us any more insight or flexibility than could be obtained by keeping to Bayesian orthodoxy? Is there any reason to suppose, for example, that choosing d to maximise $L(d,\sigma_a^2)$ is any better than letting it be a hyperparameter of the prior distribution for a, itself having a suitably woolly distribution? The latter gives you all the advantages of coherence and allows the data to dictate automatically what are the likely values of d and to give a suitably weighted posterior distribution for a.

ii) Does the gain in common sense outweigh the ad-hockery that is immediately needed as soon as coherence is abandoned? Why, for example, do we take $c_0 = 0$ in example (*a*), why the particular choice of D in examples (*b*) and (*c*), and so on? If we are not very careful we shall find ourselves in just as muddled a state as the poor frequentists.

Finally, I am puzzled by the last example on polynomial fitting where no mention is made of the purpose. Do we just want a good fit or a good prediction, or do we really want to know the "true" order of the polynomial and to estimate its coefficients? Without any context I can't judge the meaning of the results presented.

Professor Dawid disarms criticism of his paper by openly admitting that much of it is not of direct interest to Bayesians. Here at least is one statement I can broadly agree with. The paper does give me one way of telling when a frequentist is being incoherent, but frequentists are so seldom coherent that this is somewhat superfluous. Those of us who enjoy explicitly exposing the incoherence of frequentist methods might find some of the results here useful, however.

In the components of variance example (5.6), it seems essential to allow $\tau^2 < 0$ in order to get all information about σ^2 concentrated in S_3. This is not as crazy as it seems and has indeed already been advocated by Nelder (1977). Since τ^2 is the *excess* of variance between rows over variance within rows, a negative value is possible but has strange implications. The correlation between a pair of observations in different rows (value of i) has to be *greater* than that between a pair in the same row. It is hard to imagine real datasets where this would happen.

I should like to ask if any of the results in this paper throw any more light on that undefined concept of "no available information about Θ in the absence of knowledge of Φ" introduced by Kalbfleisch and Sprott (1970). I remember the concept coming under heavy attack at that time, and the authors trying hard to make it rigorous, but I cannot recall seeing any further published work.

Finally, the distinction between parameters of interest and nuisance parameters is not always at all clear. In model discrimination problems, for example, we do not know which parameters will be of interest until we have decided which model is most likely to be true. Perhaps we need to introduce the idea of nuisance models here.

D. PEÑA (*Escuela de Organización Industrial, Madrid*):

My comments on the papers for this session will be limited to the paper by Professor Akaike, because it appears to me to be the most ambitious and most polemical of the two papers, at least within the context of this conference, and because it touches areas that are more related to my particular interests and competence.

Briefly, the paper by Professor Dawid appears to me to confirm what Bayesian Statisticians already know: namely, that the treatment of nuisance parameters within the Bayesian framework is general and coherent, in contrast with the many partial solutions adopted by classical statisticians.

My criticisms of the paper by Akaike fall into three categories: (1) I do not agree with a number of the general methodological comments made in the paper; (2) I am not convinced that the goodness-of-fit criteria, based on the Kullback-Leibler measure of information, suggested by Akaike, provide a significant improvement over previously existing criteria; (3) it appears to me that the general linear model, developed by Akaike in this paper, is mainly designed to solve the problem of fitting many parameters to few observations, and therefore focuses on the solution of problems in practical statistical analysis that are, initially, so ill-defined that the investigator, no matter what methodology he uses, can learn little from the data.

Beginning with the first point, general methodological questions, I do not share the opinion, expressed by Akaike (Section 1), that Bayes procedures represent only "one possible way of utilizing the information provided by the likelihood function". I would agree, with Jeffreys and others, that Bayes methodology embodies the scientific principle of "learning from experience" in an essencially non-deterministic world. The justification of Bayesian methodology is, in my view, that it provides a unified and internally consistent approach to dealing with uncertainty, both in the context of statistical inference and decision.

Professor Akaike presents two objections to the subjective interpretation of Bayesian procedures. First, he objects to the postulate of linear ordering of preferences in Savage's axiom system, and offers an example of a preference structure that appears, at first sight, to be sensible, but in fact is not transitive. It seems clear to me that the transitivity axiom is needed in any coherent theory of decision that is to be applied to real life problems with any degree of success. Raiffa (1968, pp. 75-86) has shown, in a very convincing way as far as I can see, how it is always possible to build a "money-pump" against the intransitive subject.

The second objection, in Akaike's words (Section 2.2) is:

> "To take the parameters (as) something prespecified and assume that the prior distribution can or should be determined independently of the data distribution constitutes a serious misconception about the inferential use of the Bayes procedure".

I certainly agree that, in principle, the data distribution should be taken into account in specifying the prior distribution in the non-informative situations typical of much of statistical inference. However, this is not a new point and, in the concrete example offered by Akaike (Bernoulli versus Pascal sampling), it is not of much practical importance; see Box and Tiao (1973, pp. 45-46). The dependence of the prior distribution on the data distribution is also present in the maximal-data-information prior distributions suggested by Zellner (1977).

In summary, with respect to general methodological questions, the "Conceptual difficulties of the subjective approach" suggested by Akaike do not seem convincing to me, and therefore, it does not seem to me to be necessary to look for new foundations for Bayesian Inference.

I now move on to a second class of comments, those related to the new goodness-of-fit criteria developed by Professor Akaike. This paper introduces a new information criterion, the *ABIC*, to select the optimal value of the constant *d* in his mathematically elegant, general linear model. In essence, this new criterion, the *ABIC*, is simply the older criterion, the *AIC*, also developed by Akaike, applied to the general linear model of this paper. These statistical criteria are based primarily on the Kullback-Leibler measure of information, but their justification, as far as statistical optimality is concerned, has remained heuristic. I am sure that a strengthening of the tie between information theory and statistics is a useful research objective, but I suspect that the particular criteria presented in this paper are equivalent, in most cases, to classical statistical test criteria. To support this view, let us consider a problem frequently treated by Professor Akaike (1974, 1976, 1978) in which the minimum *AIC* is applied: The selection of the order of a stationary normal autorregressive stochastic process. In this case a model with $p + k$ parameters is chosen over a model with only p if:

$$AIC\,(p + k) < AIC\,(p).$$

The above inequality is equivalent to:

$$N\,\mathrm{Ln}\,\hat{\sigma}^2\,(p + k) + 2\,(p + k) < N\,Ln\,\hat{\sigma}^2\,(p) + 2p,$$

where $\hat{\sigma}^2\,(p + k)$ and $\hat{\sigma}^2\,(p)$ are the estimated residual variances of the two models, and N is the number of observations. This implies:

$$\mathrm{N\,Ln}\,\frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p+k)} > 2k$$

and, using the fact that $Ln\,(1 + x) \approx x$ when x is small, this reduces to:

$$\frac{N\,\{\hat{\sigma}^2\,(p) - \hat{\sigma}^2\,(p + k)\}}{\hat{\sigma}^2\,(p + k)} > 2k$$

which is equivalent to:

$$F_{K,N} = \frac{N\{\hat{\sigma}^2\,(p) - \hat{\sigma}^2\,(p+k)\}}{K\,\hat{\sigma}^2(p+k)} > 2$$

where $F_{K,N}$ is the statistic F with K and N degrees of freedom. In this calculation I have used the assumption that N is large enough so that $N - p \approx N$. Asymptotically, we obtain the classical likelihood ratio test based on the χ^2 with k degrees of freedom (Bartlett (1978), pp. 306-307):

To summarize, if N is large, the *AIC* is equivalent to a likelihood ratio test based in the χ^2 with k degrees of freedom and critical value of $2k$. The fact that this critical value remains equal to $2k$ explains the observed behaviour of the *AIC* and, in particular, its bias toward overparametrization pointed out by Shibata (1976).

My third category of comments referrs to the importance of choosing a parametrization that facilitates the process of learning from the data. To illustrate the usefulness of his general linear model, Professor Akaike considers the decomposition of a time series into trend, seasonal and irregular components, a problem that I find particularly important for those of us who are working in practical time series analysis. In this application the formulation by Akaike involves $2N$ parameters for the decomposition of N time series observations. The determination of this very large number of parameters is based mainly on a priori restrictions. This procedure can be regarded, in Bayesian terms, as equivalent to the application of a highly informative prior distribution about the structure of the decomposition.

We would expect this procedure to yield reasonable results in those cases in which seasonal structure is very clear at the outset, as it is in the cases presented in the paper. However, in cases in which the seasonal structure is not at all clear from the outset, we will face one of the following two unpromising choices: (1) to apply the same kind of restriction used by Akaike in the cases of the paper, which may well permit us to learn little from the data; or (2) to formulate a new set of restrictions with no guidelines for this selection.

For these reasons, I feel uncomfortable with this solution to the decomposition problem. I believe that learning from experience means, among other things, to allow the data to correct our a priori beliefs. To achieve this end, I would prefer procedures more in the spirit of Box-Jenkins (1970), that is, the use of a well-designed system of diagnostic checks, together with an iterative process of model-building which, of course, must place major emphasis on parsimonious parametrizations. In this sense, it seems to me that work along the lines of Box, Hillmer and Tiao (1976) is more promising than that presented in this paper for the time series decomposition problem.

In closing, I would like to thank both authors for their contributions to this session. To Professor Dawid I would like to express my regrets that my fields of interest and competence have not permited me to pay more attention to his paper, and I would like to thank Professor Akaike for the very stimulating and polemical paper that he has offered us on this occasion.

J.M. DICKEY (*University College Wales*):

The emphasis in the paper by Professor Akaike seems wrong to me. He writes, "It is almost trivial to see that no practically useful Bayes procedure is defined without the use of the likelihood function". This may be true in the narrow technical sense of the word "procedure", namely for an act-valued function defined on a sample space. However, such Bayesian methods can also be viewed as forming a mere subarea of subjective probability modeling in which expert opinion is quantified in its various complexities, joint dependencies, and conditioning on concomitant variables and on experimental data. It is then a rather special case to have statistical data on which to condition, and an imbedded statistical model, by which Bayes' theorem would be the form taken for the probability conditioning. This seems to be the view taken by De Finetti in his work, and it also describes the standpoint of my own paper in these Proceedings.

Think of the problem of probabilistically quantifying a physician's opinion of a cancer patient's survival under treatment with a combination of radiotherapy and a particular new drug. Suppose no proper statistical data is yet available. Probabilistic predictions (previsions) are needed for various types of patients and various treatment schedules. Perhaps, an experiment needs planning. How shall the expert's opinion be used now in planning the experiment and treating those patients who cannot wait for the definitive data?

Should it be used in the form of a subjective probability model fitted to elicited aspects of his opinion? Or should chaos reign in the deliberate rejection of any theory? Does no data mean nothing can yet be done? Perhaps one prefers to use subjective probability rather than have chaos. It is easy to say "yes" here to subjective probability modeling in an absence of statistical data, because there are no competing methods.

But now, if one says, "Yes, I shall quantify opinion probabilistically", what I say is, "Suppose one has a *little* bit of statistical data; does one now use some entirely different approach not based on subjective probability?" Suppose there is not enough data for maximum likelihood or for the use of an ignorance prior to yield sensible probabilistic previsions. And now I ask "What if one has a little larger amount of statistical data?"

You see what I am driving at. At what point does one throw away the notion of quantifying opinion by probability? At what point does one say, "I am no longer willing to specify a prior distribution as an expression of opinion"?.

S. GEISSER (*University of Minnesota*):

I am highly sympathetic to the view advocated by Professor Akaike and others that in certain contexts the prior distribution of a parameter need not be determined independently of the data distribution (likelihood). Whenever (as often is the case) the parameter is a hypothetical construct, unobservable, and artificially devised to promote a convenient model and useful only inasmuch as predictive distributions can be calculated, there seems to me no grave difficulty in taking this view. Professor Akaike, however, has really taken the bull by the horns when he chooses a coin tossing experiment to illustrate his view that it is irrational to adopt one and the same prior for the two sampling plans that led to the same likelihood for the parameter θ. In my view it is very difficult, if not impossible, to argue that it is rational not to adopt the same prior in this particular situation.

In this situation, if anywhere, θ comes closer to being a physical property of the coin than in most other experiments statisticians deal with. The sampling plan can in no way affect this property. Hence one can rightly argue that the two different sampling rules invoked are irrelevant towards inferring about this "physical entity". If one takes the view, as I do, that even in this case the predictive distribution of a future observation is paramount rather than the posterior distribution of θ - neither should be affected by the sampling rule once the sample is in hand. If one takes this from the ususal parametric framework (for a predictivist it is always more confortable to be able to frame the problem in terms of observables) and one can do so to a degree in this case, we can sharpen the divergence of opinion on what is rational. To my mind, there is always fuzziness in frameworks involving hypothetical unobservables. Jeffreys (1939) discusses the case where there are N binary trials with an unknown number R of one type and N-R of the other. A sample of n is drawn and T of one type observed and the predictive distribution of R obtained,assuming all possibilities are, a priori, equally likely for R. Here the sampling is hypergeometric. One could have also sampled until T was observed and hence obtained a negative hypergeometric sampling distribution for the total sampled. The "likelihood" (actually in either case it is a probability conditional on the potential observable, R) of R is unaltered as in the parametrized negative binomial - binomial situation.

It appears that here in the completely observable situation, Professor Akaike would be on very precarious ground in sustaining his view that it is irrational for the same statistician to have a single prior for R given only that the sampling plan was at issue.

With respect to Professor Dawid's paper, if one restricts one's attention to the prediction of observables or potential observables, then the problem of nuisance parameters, with its imposing glossary of terms, completely vanishes. Although this is my philosophical stance, I admit to harboring some genuine regret as to having my view universally adopted since it would preclude the appearance of much elegant research such as Professor Dawid's and many of those listed in his references.

D.V. LINDLEY (*University College London*):

The criticism of the axioms offered by Professor Akaike fails to distinguish between the descriptive and the prescriptive views. A person who has preferences like the young boy would lose money for sure and, although it may be an accurate description, it is hardly a prescription for sensible behaviour. My description of Akaike is closely related to that of a prescriptive person: he obtains sound answers for wrong reasons.

An alternative approach to polynomial fitting is available by Young (1977). He fits polynomials of very high degrees using a prior that reflects scientific opinion that low-degree polynomials are more reasonable than those of high degree. This approach finishes up with a low-degree polynomial and avoids the difficulties of choice between models. Generally, it often seems sensible for a Bayesian to fit the largest model he can. Model choice is really a decision problem of what variables to observe in a future experiment.

A. O'HAGAN (*University of Warwick*):

I find myself in disagreement with some of the things Professor Akaike has to say, for instance the whole of sections 2 and 3. But Professor Akaike has too much experience with data to produce silly analysis however misguided his philosophy might be, so I was not surprised that the technique he advocates in section 5 for estimating the variance parameters σ^2 and d is perfectly sensible. In fact, in O'Hagan (1976) I reached a similar conclusion, that one should (*a*) estimate variance parameters by the mode of their marginal distribution (after integrating out the other parameters), then (*b*) estimate the other parameters by the mode of their conditional distribution given that the variance parameters have the values obtained in (*a*). Professor Akaike does not put priors on σ^2 and d, so his step (*a*) is a maximization of "marginal likelihood".

A.F.M. SMITH (*University of Nottingham*):

The examples presented in Section 4 of Professor Akaike's paper are interesting examples of what *I* would call, in constrast to the opening paragraph of that section, "the common-sense approach to constrained least squares". If the author is interested in "the common-sense approach to Bayesian Statistics" he might try Lindley and Smith (1972).

REPLY TO THE DISCUSSION

AKAIKE, H. (*The Institute of Statistical Mathematics, Tokyo*):

Just before the presentation of my paper I felt that I was rather out of place. After receiving the comments I recognized that my participation in the meeting was extremely rewarding. I must express my sincere thanks to the organizing committee and those who contributed to the discussion for providing me such an enjoyable intellectual experience.

Professor Barnard disagrees with my critical view of Savage's postulate on linear ordering of preference. Nevertheless, by the recent review article of Professor Good (1979), it seems that Savage himself considered his system of subjective probability incomplete, as it rejects the concept of randomization. To accept the concept of randomization is equivalent to accepting the impossibility of uniquely specifying a prior distribution.

Professor Barnard's warning against assuming ignorance without sufficient analysis of a particular situation is extremely valuable. My recent experience on developing a smoothness prior for the distributed lag model treated by Shiller shows that a Bayesian model can produce a significantly distorted image of the reality (Akaike, 1979). It seems that the only sensible way out of this difficulty is to develop several alternative Bayesian models and evaluate their likelihoods with respect to the available data.

Professor Barnard's general opinion on the use of Bayesian models is so close to mine that it is almost impossible for me to point out any significant differences. The basic idea here is to base the justification of the use of a Bayesian model on the following identity

$$\text{objective} = \text{social} = \text{long run.}$$

We consider that the information expressed in terms of a prior distribution must at least be communicable. This communicability can only be gained by placing the prior distribution within the context of its particular application. This observation, I think, is the gist of Professor Barnard's comments.

Finally, I wholeheartedly support Professor Barnard's view on the danger of the excessive separation of doctrines of statistics. Each doctrine tends to suppress activities outside of it. At one point this tendency begins to act against the progress of human knowledge. A real innovation can never be placed properly within an existing doctrine and there should be no end of the progress of human knowledge.

Professor Freeman surprises me by rejecting the basic Bayesian principle of rationality, the maximization of expected utility. He then violates the teaching of subjective probability by ignoring, without reason, the information of whether the sampling is direct or inverse in the case of a binomial experiment.

Professor Freeman is particularly sensitive to "objectivity", as a sensible statistician should always be. Statistics always deals with data which represent the outside world. Even if the choice of a data distribution is subjective, the likelihood determined by data is an objective evaluation of the assumed data distribution. Even Shakespeare cannot fight against the objectivity of data.

The prudence shown by Professor Freeman against the numerical results reported in my paper is impressive. Particularly his preference of real examples to artificial ones reveals his position to consider statistics as something related with the outside world.

To the two questions raised by Professor Freeman I answer as follows: (i) The idea stressed by the examples discussed in the paper is the importance of the technical understandability of prior distributions. The examples also suggest the utility of defining an objective procedure of the choice of a prior distribution. Any subjectively chosen proper prior distribution, however wooly it may be, cannot be free from a possible gross misspecification. (ii) There should be no problem in choosing c_0 and D, if their technical meanings are clearly understood.

As to the predicament of Professor Freeman about the last example on polynomial fitting my explanation is that I am only interested in getting a good predictive distribution. There is no meaning in talking about the "true" order, as this is infinite.

Dr. Peña considers that the conceptual difficulties of the subjective approach is not substantial. His conclusion is based on two observations. The first is that there are Bayesians, like Jeffreys, Box, Tiao and Zellner, who treat the problem of inference to Dr. Peña's satisfaction. But these people are not subjective Bayesians. They all accept the use of improper prior distributions, which is unacceptable to strictly subjective Bayesians. Dr. Peña's second observation is that my criticism of Savage's postulate of linear ordering of preference is already sufficienty disproved by Raiffa's "money-pump" argument. Raiffa's explanation starts by assuming that a person with incoherent preference has made a decision. What I am insisting with the example of the boy with the preference described in the text of my paper is that he is trapped in a state of indecision. Thus Raiffa's "money-pump" argument does not constitute any disproof of my criticism of the difficulty of Savage's axiom.

As to the criticism of Dr. Peña of the information criterion I must say that the classical tests are often disguised realizations of estimations when there are several possible models. The optimality of the minimum *AIC* procedure is discussed by Akaike (1978b) and Shibata (1980), but what I am interested in here is the use of the concept of likelihood or entropy in Bayesian modeling rather than the use of minimum *AIC* type procedure.

Dr. Peña's criticism of the use of the general linear model for seasonal adjustment suprises me. The whole procedure is objectively defined. It is simple and can be tested by anyone who is interested in it. The procedure is completely free from the *ad hoc* manipulations of data, at the beginning and end of the time series, by Census Methods of seasonal adjustment. I do not deny the possibility of other procedures, but I must mention that there is nothing like a canonical form for a system varying with time and that this makes the ordinary parametric approach to the seasonal adjustment problem very difficult. The main point of the introduction of the present general linear model is the clarification of the importance of technical understandability of a prior distribution. I hope that Dr. Peña would agree with me to consider the fact that a computer program is already in existence and is producing useful outputs without much human intervention as a clear demonstration of the power of this approach.

Professor Lindley considers my criticism of Savage's axiom to be due to the confusion of descriptive and prescriptive views. My criticism of subjective Bayesians is that their prescriptive attitude looks very much like the attitude of a physician who gives a huge collection of precriptions of drugs to a patient and leaves the burden of identifying the proper choice to the patient. The "money-pump" argument tells the patient that he must take a drug described by the physician but does not help him in making his choice.

Young's (1977) paper on polynomial fitting is not free from the basic difficulty. The prior distribution contains two hyperparameters. Apparently Young did not propose any systematic approach to the choice of the hyperparameters.

Dr. O'Hagan tells me that I am producing sensible result with the help of a misguided philosophy. In his 1976 paper, Dr. O'Hagan makes use of an improper prior distribution. The result mentioned in his comment is then obtained by adjusting the Bayesian model so as to produce a result consistent with the result obtained by conventional statistics. These observations show that he himself is subscribing to the "misguided" philosophy, the common-sense approach to statistics.

Professor Smith reminds me that the paper by Lindley and Smith (1972) is a pioneering work on the common-sense approach to Bayesian statistics. Actually Lindley and Smith accept the use of an improper prior distribution, which is not acceptable to strict Bayesians. The paper demonstrates the point that the technical understandability of the prior distribution is the key to the successful application of a Bayesian model. Certainly, this is one of the themes of my present paper, but my main emphasis is on the use of likelihood as an objective measure of the goodness of a model. Even the goodness of a Bayesian model can be checked by comparing the likelihoods of competing models.

Professor Geisser is sympathetic to my common-sense approach but he fears, with Professor Hill at the time of the meeting, that I am touching on a too delicate subject when I referred to the direct and inverse binomial experiments. It looks to me that he is too much influenced by the so-called objective theory of probability. Within the statistical context, it must be accepted, every probability is conditional on available information. If we knew whether the experiment was direct or inverse, this constitutes a part of our prior information. Thus the assumption of prior independence of the probability of head in a coin tossing with the information of the type of experiment is acceptable only under certain specific circumstances.

Consider the situation where you are served a piece of pie. When you know that the pie was prepared by a cook who is notorious for poisoning your attitude towards the pie will be different from that when you know that the cook had a perfect record. Dr. Peña drew my attention to Box and Tiao (1973, pp. 45-46) who accepted the difference of the ignorance priors for the two sampling schemes. Thus I am not alone here.

Professor Dickey points out that my emphasis on likelihood is wrong and reminds me of the importance of interpreting a prior distribution as an expression of a personal opinion. In Professor Dickey's argument I sense, as in almost every argument by subjectivist Bayesians, a rash inclination towards the assumption of the state of ignorance, or of no information. I consider this a dangerous sign. Particularly, when Professor Dickey forcefully puts forward the dichotomy between subjective probability and chaos, I see a curious analogy between his position and that of epistemological traditionalism observed by Popper (1965, p.6) who states 'we can interpret tradicionalism as the belief that, in the absence of an objective and discernible truth, we are faced with the choice between accepting the authority of tradition, and chaos'.

We notice that Professor Dickey's argument gains weight only when he uses the word "expert opinion" instead of an arbitrary "opinion". What discriminates an expert's opinion from a layman's is that the former is backed by experiences, either of the expert's own or someone else's. The experience are appreciated only when they constitute objective information. In constructing his prior distribution, the expert will evaluate, at least informally, the likelihoods of various conditional statements with respect to this information. It is the objectivity thus obtained that makes an expert's prior distribution respectable.

Now we come to the discussion of the state of ignorance. For a person who tries to collect information to establish a hard prior opinion, it is a rule rather than exception that he faces the lack of information which prevents him from determining a unique prior distribution. The impossibility of uniquely determining his prior distribution, typycally represented by the introduction of hyperparameters, is the representation of the lack of information and however hard he may try to elicit the details of his opinion he cannot produce information out of nothing. Nevertheless, due to the limitation of time, he has to make a decision, and this requires a unique choice of a prior distribution. How should he act in such a situation? The answer seems clear. The effort in defining a prior distribution is mainly directed towards delineating relatively important possibilities. When the effort comes to a halt due to the lack of information, we come to the phase of making a decision. The situation is typically represented by that of planning the experiment in Professor Dickey's comment. Here the emphasis is on paying attention to every possibility. Who will favor a physician's whim to a carefully designed experiment which takes into account every possible course of patient's condition? Thus, at the point where the collection of further relevant information becomes impossible, the emphasis is switched from restraining to dispersing the distribution of the prior probability. This may sometimes lead to the use of improper prior distributions. The effect of this dispersing is evaluated by its effect on the resulting predictive distribution. Here the recognition of the necessity of switching the point of view, during the process of developing a prior distribution, seems crucial.

The above is an amplified version of the procedure for the construction of a prior distribution discussed in my paper. A simple but concrete example of application of this procedure is discussed in Akaike (1980).

Thus in our approach the subjective elements are always exposed to some objective tests through the use of prior experiences or data, and the somewhat obscure concept "opinion", required to complete a prior distribution, is replaced by a description of a strategy for making a decision. This strategy and its design principle are described objectively and can be tested in the long run through the accumulation of experiences of its use by a scientific community. Thus, contrary to the suggestion of Professor Dickey, we do not put much emphasis on the interpretation of a prior distribution as an expression of "opinion".

DAWID, A.P. (*The City University, London*):

Should the Bayesian be interested in concepts springing from a frequentist or "prespecification" approach to inference, or can he afford to dismiss them cursorily as "incoherent"? Although I am fully committed to the Bayesian position, I can't accept that the only good ideas are those had by Bayesians. Consequently I regard it as practically important, as well as theoretically amusing, to investigate non-Bayesian ideas, and find out how they relate to Bayesian ones. So perhaps I should revoke my suggestion that some of the definitions of my paper are of no interest to Bayesians, for if we are to be good statisticians (which is surely more important than being coherent) we must not dismiss such concepts out of hand — at any rate, not before a thorough investigation of the type I have attempted.

While agreeing with Professor Barnard that we could well drop the term "nuisance parameter", I am puzzled by his suggestion that we are guilty of some sort of sin if we lay down, before getting data, that we are only interested in what we can learn about the parameter Θ. If this is an example of "prespecification", I can only conclude that there must be a large overlap between that approach and Bayesian ideas. If we are faced with a decision problem in which the parameter enters the loss function only through Θ, why should we not make an inference about Θ alone, whatever the data may turn out to be, and whatever the dependence between Θ and some nuisance parameter Φ?

I do, however, accept Professor Geisser's point that the formulation of the problem in terms of parameters at all may be mistaken. A reformulation involving the unknown values of future observations would involve quite different theory, at least for the frequentist. Such an approach might perhaps be used in the model discrimination context mentioned by Professor Freeman, since, while we may not know what parameters are of interest, we will surely be able to pinpoint what it is that we should like to be able to predict. Nevertheless, for the Bayesian, an emphasis on prediction is not a pre-requisite for the problem of nuisance parameters to disappear - he never had a problem in the first place. It is classical ideas which present problems. If we take a predictive standpoint, then it becomes appropriate to compare the straightforward Bayesian approach to prediction with classical counterparts (see, for example, Section 6 of Dawid, 1979a). But that is another story.

Barnard's discussion of a "model-adjustment" parameter is important. It attacks the problem of the robustness of an inference about the parameter of interest. His assumption is that the likelihood, while not being a function of θ only, is nevertheless approximately so, for most data. If the data suggests that this approximation is good, then we can pretty well ignore the fact that there are really some nuisance parameters around. If, however, we have exceptional data, we may have to be more careful. This suggests an interesting line of research; in particular, how would the Bayesian formalize the property that, to a good approximation, his model involves only the parameter of interest? This kind of problem, in which approximations may be valid for some data values, but not for others, is of great general importance to a sensible Bayesian approach. In particular, the marginalization paradox does not rule against using the paradoxical posterior distribution for the data at hand, but warns that it cannot be a good approximation to a coherent posterior for *all* possible data values. In concerning ourselves with these things, we are, of course, leaving the pre-specification approach squarely behind, and rightly so.

The incoherence in Example (5.6) is not really concerned with the question whether or not we can have $\tau^2 < 0$, as suggested by Professor Freeman. As I point out, we could, for example, get variation-independence between σ^2 and (μ, σ_0^2) if σ^2 and τ^2 are subject to $\sigma^2 \leq 1$, $\tau^2 \geq (1-\sigma^2)/J$. The real difficulty is the dependence on J. Thus, while one might argue that it is coherent to use only S for inference about σ^2 for the experiment performed, one could not allow this same argument simultaneously for another such experiment, with different J. This is analogous to the discussion at the end of the previous paragraph, with the difference that, there, we had to worry about inferences from different data in one experiment, while here we must worry about different experiments. But the comparison of different experiments is a valid and important concern of the theory of coherence.

As for the concepts of "no available information about Θ in the absence of knowledge of Φ", the various ideas of *S, G, M*-ancillarity etc., all express classical attempts to capture this notion: I refer Professor Freeman to Barndorff-Nielsen's book. I don't think there is a full-blooded Bayesian interpretation, because of the difficulty of defining "absence of knowledge". Marginal ancillarity is not really appropriate, depending as it does very much on the form of prior knowledge about Φ. But if we once again drop a pre-specification approach, it may be that the concept can be given some meaning in terms of robustness or approximation, relevant only for certain data and classes of prior distributions.

REFERENCES IN THE DISCUSSION

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-722.

— (1976). Canonical correlation analysis of time series and the use of information criterion. In *System Identification: Advances and Case Studies*. (Mehra and Lainiotis eds.) New York: Academic Press.

— (1978). On the likelihood of a time series model. *The Statistician*, **27**, 217-235.

— (1979). Smoothness priors and the distributed lag estimator. *Tech. Report No. 40*, Stanford University.

BARTLETT, M.S. (1978). *An introduction to Stochastic Processes*. Cambridge: University Press.

BOX, G.E.P. and JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control*. New York: Holden-Day.

BOX, G.E.P. and TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis*. New York: Addison Wesley.

BOX, G.E.P., HILLMER, G.C. and TIAO, G.C. (1976). Analysis and modelling of seasonal Time Series. Presented at *NBER/Bureau of the Census Conference on Seasonal Analysis of Economic Time Series*. Washington, D.C.

GOOD, I.J. (1979). Book review of *Logic, Law and Life: Some Philosophical Complications*, (R.G. Colodomy ed.) *J. Amer. Statist. Assoc.* **74**, 501-502.

KALBFLEISCH, J.D. and SPROTT, D.A. (1970). Application of likelihood method to models involving large numbers of parameters (with discussion). *J. Roy. Statist. Soc. B.* **32**, 175-208.

LINDLEY, D.V. and SMITH A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy Statist. Soc. B.* **34**, 1-41.

NELDER, J.A. (1977). A reformulation of linear models (with discussion). *J. Roy. Statist. Soc. A* **140**, 48-77.

O'HAGAN, A. (1976). On posterior joint and marginal modes. *Biometrika* **63**, 329-333.

POPPER, K.R. (1965). *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books. Also in Harper and Row, (1968) New York.

RAIFFA, H. (1968). *Decision Analysis*. New York: Addison-Wesley.

SHAKESPEARE, W. (1598). *The Second Part of the History of Henry the Fourth, I. 3*, 35-36. London: Wise and Aspley.

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117-126.

— (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147-164.

YOUNG, A.S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika* **64**, 309--317.

ZELLNER, A. (1977). Maximal data information prior distributions. In *New Developments in the Applications of Bayesian Methods*. (A. Aykas & C. Brumat eds.) Amsterdam: North-Holland.

Least squares approximation in Bayesian analysis

MICHEL MOUCHART* and LÉOPOLD SIMAR**

*Université Catholique de Louvain

**Facultes Universitaires Saint Louis-Bruxelles

SUMMARY

The paper presents in a simple and unified framework the Least-Squares approximation of posterior expectations. Particular structures of the sampling process and of the prior distribution are used to organize and to generalize previous results. The two basic structures are obtained by considering unbiased estimators and exchangeable processes. These ideas are applied to the estimation of the mean. Sufficient reduction of the data is analysed when only the Least-Squares approximation is involved.

Keywords: LINEAR BAYES, LEAST SQUARES, CREDIBILITY THEORY

1. INTRODUCTION

1.1. *General Formulation*

Consider a random vector (θ',x') with $\theta \in R^q$ and $x \in R^p$. In what follows, x will typically represent (functions of) observations and θ will represent either (functions of) parameters or future observations. In all of this paper, (θ',x') is assumed to be square-integrable.

In Bayesian analysis, attention is often directed toward computing the posterior expectation $E(\theta|x)$. This is, *e.g.,* the Bayesian decision rule under quadratic loss. In this paper we consider simple (*i.e.* linear) approximations of $E(\theta|x)$; they will be denoted $\hat{E}(\theta|x)$. Under the Least-Squares (L.S.) criterion, the best linear approximation of $E(\theta|x)$ is also the best linear approximation of θ. In this second interpretation, $\hat{E}(\theta|x)$ may be viewed as a "best linear estimator of θ". Doob (1953) suggests that $E(\theta|x)$ be called the best L.S. approximation of θ and $\hat{E}(\theta|x)$ the wide-sense version of $E(\theta|x)$.

In order to write explicitly $\hat{E}(\theta|x)$, we partition the vector of the expectations and the variance-covariance matrix in the following way:

Trab. Estadist. (1980)
Vol. 31 pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 205-246 Valencia, Spain: University Press.

4. Approximations

INVITED PAPERS

MOUCHART, M. (*Université Catholique de Louvain*)
and
SIMAR, L. (*Facultés Universitaires Saint-Louis, Buxelles*)

Least squares approximation in Bayesian analysis

LINDLEY, D.V. (*University College London*)

Approximate Bayesian methods

DISCUSSANTS

BROWN P.J. (*Imperial College, London*)
GOLDSTEIN, M. (*University of Hull*)
BERNARDO, J.M. (*Universidad de Valencia*)
DAWID, A.P. (*The City University*)
FRENCH, S. (*University of Manchester*)
GOOD, I.J. (*Virginia Polytechnic and State University*)
GREN, J. (*Econometric Institute, Warsaw*)
O'HAGAN, A. (*University of Warwick*)

REPLY TO THE DISCUSSION

$$E\begin{pmatrix}\theta\\ x\end{pmatrix} = \begin{pmatrix}E(\theta)\\ E(x)\end{pmatrix} = \begin{pmatrix}m\\ E(x)\end{pmatrix} \tag{1.1}$$

$$V\begin{pmatrix}\theta\\ x\end{pmatrix} = \begin{pmatrix}V_{\theta\theta} & V_{\theta x}\\ V_{x\theta} & V_{xx}\end{pmatrix} \tag{1.2}$$

Under the Least-Squares criterion, the best linear approximation of $E(\theta\,|\,x)$ is known to be:

$$\hat{E}(\theta\,|\,x) = m + V_{\theta x}\, V_{xx}^{-1}(x-E(x)). \tag{1.3}$$

It is important to point out that formula (1.3) is valid irrespective of the form of the distribution of (θ,x); the only restriction being the existence of second-order moments. Reading (1.3) component-wise, we conclude that each component of $\hat{E}(\theta\,|\,x)$ is also the L.S. approximation of the corresponding component of θ (or of $E(\theta\,|\,x)$); more formally:

$$\hat{E}(\theta\,|\,x) = \begin{pmatrix}\hat{E}(\theta_1|x)\\ \hat{E}(\theta_2|x)\end{pmatrix} \tag{1.4}$$

Formula (1.3) is computationally simple and needs only the specification of the first two moments; this gives it some properties of robustness.

When θ represents parameters of the sampling distribution, this specification will often rely on the following decomposition, based on averaging over sampling moments:

$$E(x) = E_\theta E(x\,|\,\theta) \tag{1.5}$$

$$V_{xx} = E_\theta V(x\,|\,\theta) + V_\theta\ E(x\,|\,\theta) \underset{\text{def}}{=} V_1 + V_0 \tag{1.6}$$

$$V_{\theta x} = E_\theta[\theta E(x'\,|\,\theta)] - E(\theta)E(x') = \operatorname{cov}\big(\theta, E(x'\,|\,\theta)\big) \tag{1.7}$$

Apart from the ease of computation and the aspect of robustness, the accuracy of the L.S. approximation is often crucial. Let us introduce

$$\eta = \theta - \hat{E}(\theta\,|\,x) = (\theta - m) - V_{\theta x}V_{xx}^{-1}\,\big(x-E(x)\big) \tag{1.8}$$

Clearly, η has zero mean and is uncorrelated with x. Often, one would like to analyse the accuracy of the approximation *for a given* x. We have the following posterior moments:

$$E(\eta\,|\,x) = E(\theta\,|\,x) - \hat{E}(\theta\,|\,x) \tag{1.9}$$

$$V(\eta\,|\,x) = V(\theta\,|\,x). \tag{1.10}$$

Unfortunately, these quantities are generally at least as difficult to compute as $E(\theta\,|\,x)$ itself. However, $V(\eta)$ is easily computed from (1.8)

$$V(\eta) = V_{\theta\theta} - V_{\theta x}V_{xx}^{-1}V_{x\theta} \tag{1.11}$$

This formula again depends only on second moments of (θ,x) (directly computable from prior and sampling moments when θ represents a parameter). We now decompose $V(\eta)$ as follows:

$$V(\eta) = E_x V(\eta\,|\,x) + V_x E(\eta\,|\,x). \tag{1.12}$$

The dispersion of η has therefore two components: by (1.10) the first one is due to the average posterior variance of θ and the second one comes, by (1.9), from the possible non-linearity of $E(\theta\,|\,x)$. Therefore $V(\eta)$ gives an upper bound for the average posterior variance of θ:

$$E_x V(\theta\,|\,x) \leq V(\eta) \tag{1.13}$$

where $\leq$ is written in the sense of positive-definite, symmetric (P.D.S) matrices, and with equality if and only if the true regression is linear $\big($*i.e.* $E(\theta\,|\,x) = \hat{E}(\theta\,|\,x)\big)$.

In particular,

$$V(\eta) = V(\theta\,|\,x) \text{ for any } x$$

if and only if the true regression of θ on x is

(i) linear $\big(E(\theta\,|\,x) = \hat{E}(\theta\,|\,x)\ \ a.s.\big)$

(ii) homoscedastic ($V(\theta\,|\,x)$ constant *a.s.*).

As this is the case for the normal distribution, one may interpret the L.S. approximation as adjusting an overall normal distribution on (θ,x) with identical first two moments. In other words, the formula used to compute $\hat{E}(\theta|x)$ (*i.e.* (1.3)) and $V(\eta)$ (*i.e.* (1.11)) may be viewed as the conditional mean and variance of that normal approximation. This feature has been demonstrated by Doob (1953 - Chap. 1) and Hartigan (1969); both suggested the terminology of "linear expectation" for $\hat{E}(\theta|x)$; Hartigan also suggested "linear variance" for $V(\eta)$, and even used the notation "$V(\theta|x)$", but this appears to be ambiguous and will not be used here.

From a decision point of view, let us consider $\hat{E}(\theta|x)$ as a decision rule. From (1.8), $V(\eta)$ appears as the Bayesian mean-squared error matrix of $\hat{E}(\theta|x)$:

$$M\,S\,E(\hat{E}(\theta|x)) \equiv E(\theta - \hat{E}(\theta|x))\;(\theta - \hat{E}(\theta|x))' = V(\eta). \qquad (1.14)$$

Under a quadratic loss associated with a decision rule $t = t(x)$:

$$\ell(t,\theta) = (t-\theta)'\,A(t-\theta) \qquad A : \text{SPDS} \qquad (1.15)$$

the Bayesian risk associated with $\hat{E}(\theta|x)$ is:

$$R(\hat{E}(\theta|x)) \equiv E\;\ell(\hat{E}(\theta|x),\theta) = tr\;A\;\;V(\eta). \qquad (1.16)$$

In any case, $V(\eta)$ will determine the decisional accuracy of $\hat{E}(\theta|x)$.

1.2. *General Comments and Objectives of the Paper*

We developed an interest in L.S. approximations when supervising a student's thesis on credibility theory (Bouchat (1977)). We then became aware that the idea of L.S. approximation to Bayesian solutions had been widely used in various fields of applications with different terminologies and striking duplication of results. It has been used since 1920 in actuarial sciences under the heading of credibility theory. An overview may be found in Bühlman (1970), de Vijlder (1975) or Kahn (1975). Recent developments are also due to Bülhman (1971) and Jewel (1974 a, b, c). Hartigan (1969) and Goldstein (1975 a, b, 1976), under the heading of linear Bayes methods, analyse the L.S. approximations in various particular statistical problems. Stone (1963) and Dickey (1969) arrive at similar methods when looking for robust Bayesian procedures.

A recurrent theme in the above literature considers whether the L.S. approximations is exact or not, *i.e.* whether or not $\hat{E}(\theta|x) = E(\theta|x)$. Bailey (1950) and Mayerson (1964) have shown that particular combinations of prior probability and likelihood yield exact credibility for the mean of a process. Jewel (1974 b, c) extended these results for the exponential family under natural-conjugate prior. Kagan, Linnik and Rao (1973, addendum B) give conditions for the linearity of Bayes estimators. Recently, Diaconis and Ylvisaker (1979) have characterized conjugate prior measures through the property of linear posterior expectation of the mean of the process. By so doing they not only extend previous results on exact L.S. approximations, but they also linked this problem to the admissibility of linear estimator under quadratic loss. (See also Kagan *et al* (1973, Chap.7).) A somewhat different approach is to characterize joint distributions (on (θ,x)) having linear expectation $E(\theta|x)$. Thus, Lukacs and Laha (1964, Chap. 6) give a necessary and sufficient condition in terms of characteristic functions.

During the revision of this paper we also became aware of recent results by Goel and DeGroot (1979) and by Goel (1979) characterizing linearity of posterior expectations in linear regression and in a scale parameter family.

This problem of exact approximation will not be pursued further in this paper. Instead our main objective is to present in a simple and unified framework previous results otherwise stated in particular contexts. By so doing, we simplify unnecessarily complicated results and remove ambiguities (which possibly induced errors).

The unifying argument is given in the general formulation of the previous section and is essentially summarized in the formulae giving $\hat{E}(\theta|x)$ and $V(\eta)$ (*i.e.* (1.3) and (1.11)). In this very simple framework, the presentation is organized according to particular structures of the first two moments (1.1) and (1.2): focusing attention on these particular structures induces natural generalizations of previous results and clarifies the role of the given assumptions. In particular, it may suggest suitable transformations (of the observations or of the parameters) in order to take advantage of specific structures (both in the prior information and in the sampling process).

Finally we systematically analyse the case of several parameters. It appears that treating each parameter individually or treating all parameters together does not affect the computation of $\hat{E}(\theta|x)$ or of the diagonal elements of $V(\eta)$. However, the role of the simultaneity in the inference shows up in the off-diagonal elements of $V(\eta)$ and so affects its inverse, associated with the concept of precision.

In Section 2, we consider two particular structures induced by the use of unbiased estimators and by some properties of exchangeability in the sampling process. It is shown that those cases provide peculiar forms of the

L.S. approximation under more general conditions than previously presented. The last section addresses itself to the question of sufficient reduction of the data when *only* the L.S. approximation is involved.

2. PARTICULAR STRUCTURES

Formula (1.3) gives a rather general framework to treat L.S. approximation in Bayesian analysis. For instance, in non-parametric situations θ may be a finite-dimensional characteristic of an infinite dimensional parameter (*viz.* the distribution function of the observation). Similarly, x may be either a full sample result or a statistic defined on a more complete sample result. Note however that the specification of $V_{\theta x}$ and V_{xx} is not always easy. It is then important to choose a suitable statistic x carefully and to take advantage of the particular structure of both the prior information and the sampling process. The object of this section is to analyse two particular structures which prove to be basic for the L.S. approximations.

2.1. *Use of Unbiased Estimator*

Suppose that we first reduce the sample to an unbiased estimator of θ:

$$E(x|\theta) = \theta. \tag{2.1}$$

Clearley, in this particular case, $p = q$. This structure implies that:

$$E(x) = E(\theta) = m \tag{2.2}$$

$$V_{\theta x} = V_{\theta\theta} = V\,E(x|\theta) = V_0. \tag{2.3}$$

Then $\hat{E}(\theta|x)$ may be written as

$$\hat{E}(\theta|x) = V_1(V_1+V_0)^{-1}m + V_0(V_1+V_0)^{-1}x \tag{2.4}$$

If V_0 and V_1 are both regular, this simplifies to

$$\hat{E}(\theta|x) = (V_0^{-1}+V_1^{-1})^{-1}\,[V_0^{-1}m + V_1^{-1}x] \tag{2.5}$$

$\hat{E}(\theta|x)$ appears as a weighted matrix average between $E(\theta)$ and x; *i.e.* $\hat{E}(\theta|x)$ has the form:

$$\hat{E}(\theta|x) = Am + (I-A)x. \tag{2.6}$$

Note that this derivation is very easy and is implied only by the property of unbiasedness. It has appeared frequently in the literature, in particular for the case $p = q = 1$ with θ being the population mean and x the sample mean. This formula is familiar for the Bayesian inference on the mean of a normal process where, in this case, $\hat{E}(\theta|x) = E(\theta|x)$ (see *e.g.* Raiffa and Schlaifer (1961)) or in credibility theory (see *e.g.* Bühlmann (1970)).

The average measure of accuracy $V(\eta)$, given in (1.11) becomes

$$V(\eta) = (V_0^{-1} + V_1^{-1})^{-1} \tag{2.7}$$

It is illuminating to write down the upper bound for the average posterior variance, in (1.13), in terms of "mean" precisions (where "mean" stands in the sense of harmonic mean, *i.e.* the inverse of the expectation of the inverse)

$$[EV(\theta|x)]^{-1} \geq [V(\theta)]^{-1} + [EV(x|\theta)]^{-1} \tag{2.8}$$

Thus the "mean" posterior precision is at least equal to the prior precision plus the "mean" sampling precision, with equality if and only if $E(\theta|x)$ is linear in x. This addition of precision is familiar (with equality) for the Bayesian inference on the mean of a normal process. In the scalar case, (2.8) has also been derived by Finucan (1971). Note however that this rule of additive precision should not be used componentwise unless $V(\theta)$ and $V(x|\theta)$ are both diagonal, which is fairly unusual.

In the light of formula (2.6) it may be illuminating to rewrite (2.7) as follows:

$$V(\eta) = A\,V_0\,A' + (I-A)\,V_1\,(I-A)' \tag{2.9}$$

This fact has been noticed by Stone (1963) for the estimation of a mean in the one-dimensional case (with x being the sample mean).

Suppose one is ready to specify the functional form of the sampling distribution but that the computation of $E(\theta|x)$ is difficult or that robustness w.r.t. the prior specification is desired. In such a case, Rao-Blackwellization may be useful. Let s be a sufficient statistic and $x^* = E(x|s,\theta) = E(x|s)$. Then $\hat{E}(\theta|x^*)$ will improve $\hat{E}(\theta|x)$ in the following sense. Let starred symbols be associated with x^* instead of x. Clearly $V_0^* = V_0$; furthermore $V_1^* \leq V_1$ by Rao-Blackwell's theorem. Therefore, from (2.7), $V(\eta^*) \leq V(\eta)$.

2.2. *Exchangeability*

2.2.1. *Introduction*

We now consider exchangeable processes, i.e. processes where the finite dimensional distributions are invariant under permutation of indices (see *e.g.* Hewitt and Savage (1955)). This class of processes generalizes the class of I.I.D. processes and also includes the mixtures of I.I.D. processes. Thus these processes arise naturally when nuisance parameters are integrated out so as to get marginalized likelihood (and prior distribution) on the parameters of interest alone. Integration of part of the parameters may also be motivated by paying attention to robustness: in a two parameter problem, for instance, the prior distribution $D\ (\theta_2|\theta_1)$ may be rather easily assigned while on θ_1 a more robust procedure may be preferred, *e.g.* by assigning only the first moment of θ_1.

Here we concentrate attention on the first two moments of a finite sequence $x = (x_1,\ldots,x_n)$ generated by such a process. In this case, $p = n$, the sample size, and q is arbitrary. We first analyse the implications of exchangeability only on the first moment, then on the first two moments: we shall call these processes first-order and second-order exchangeable.

These processes will give characterization of L.S. approximations similar to (2.4) and (2.5).

2.2.2. *First-order exchangeability*

For expository purposes, it is convenient, and not restrictive, to specify the first component of θ as the sampling expectation of the process. First-order exchangeability is then characterized by

$$E(x|\theta) = \theta_1\,\mathbf{1} \tag{2.10}$$

where $\mathbf{1} = (1\ \ 1\ldots1)' \in \mathrm{R}^n$.

Let us decompose $E(\theta)$ and $V(\theta)$ as follows:

$$E(\theta) = [m_i] \qquad i = 1,\ldots,q \tag{2.11}$$

$$V_{\theta\theta} = [v_{ij}] = [v_1\ldots v_q] \qquad i,j = 1,\ldots,q \tag{2.12}$$

where v_i is the *i-th* column of $V_{\theta\theta}$. First order exchangeability implies

$$E(x) = m_1\mathbf{1} \tag{2.13}$$

$$V_{\theta x} = v_1\mathbf{1}' \tag{2.14}$$

$$V_{xx} = V_1 + v_{11}\mathbf{1}\mathbf{1}' \tag{2.15}$$

The L.S. approximation now becomes

$$\hat{\mathrm{E}}(\theta|x) = m + [1 + v_{11}\mathbf{1}'V_1^{-1}\mathbf{1}]^{-1}\ \ v_1\mathbf{1}'V_1^{-1}(x-m_1\mathbf{1}) \tag{2.16}$$

and the average measure of precision (1.11) becomes

$$\begin{aligned} V(\eta) &= V_{\theta\theta} - [1 + v_{11}\mathbf{1}'V_1^{-1}\ \mathbf{1}]^{-1}\ \mathbf{1}'V_1^{-1}\mathbf{1}\ v_1v_1' \\ &= V_{\theta\theta} - [v_{11} + (\mathbf{1}'V_1^{-1}\mathbf{1})^{-1}]^{-1}v_1v_1' \end{aligned} \tag{2.17}$$

This involves a rather peculiar rule of additive precision analogue to (2.8) (for details, see appendix):

$$[EV(\theta|x)]^{-1} \ \geq\ V^{-1}{}_{\theta\theta} + \mathbf{1}'V_1^{-1}\mathbf{1}\ \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \tag{2.18}$$

with equality if and only if $E(\theta|x)$ is linear in x.

Note that for the (harmonic) mean of the posterior precisions the sampling improves the lower bound of the element corresponding to θ_1, the mean of the process, only and for θ_1 this improvement is given by the element (1,1) of (2.18):

$$\left[E\ V(\theta|x)\right]^{-1}_{11} \geq (v_{11}-v_{12}V_{22}^{-1}v_{21})^{-1} + \mathbf{1}'V_1^{-1}\mathbf{1} \tag{2.19}$$

where $V_{\theta\theta}$ has been partitioned as follows:

$$V_{\theta\theta} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & V_{22} \end{bmatrix} \tag{2.20}$$

If, from the start, the model had been marginalized on θ_1, formulae (2.16) and (2.17) would have given:

$$\hat{\mathrm{E}}(\theta_1|x) = [v_{11}^{-1} + \mathbf{1}'V_1^{-1}]^{-1}[(m/v_{11}) + \mathbf{1}'V_1^{-1}x] \tag{2.21}$$

$$V(\eta_1) = [v_{11}^{-1} + \mathbf{1}'V_1^{-1}\mathbf{1}]^{-1} \tag{2.22}$$

The role of the simultaneity of the θ_i's may be appreciated by comparing the inverse of (2.22) and the expression (2.19): they are equivalent if, a priori, θ_1 is uncorrelated with the other θ_i's (*i.e.* $v_{12} = 0$).

2.2.3. *Second-order exchangeability*

As for first-order exchangeability, we specify the first three components of θ as follows:

$$\theta_1 = E(x_i|\theta) \qquad i=1,\dots,n \tag{2.23}$$

$$\theta_2 = V(x_i|\theta) \qquad i=1,\dots,n \tag{2.24}$$

$$\theta_3 = \operatorname{cov}(x_i,x_j|\theta) \qquad i,j = 1,\dots,n \quad i\neq j. \tag{2.25}$$

Second-order exchangeability is characterized by the following two conditions:

$$\mathrm{E}(x|\theta) = \theta_1\mathbf{1} \tag{2.26}$$

$$V(x|\theta) = (\theta_2\text{-}\theta_3)I_{(n)}+\theta_3\mathbf{1}\mathbf{1}' \tag{2.27}$$

where, again, $\mathbf{1} = (1,1,\dots1)' \in \mathbf{R}^n$ and (θ_2,θ_3) are restricted by:

$$(-\theta_2/n\text{-}1)<\theta_3<\theta_2. \tag{2.28}$$

Like $V(x|\theta)$, $V_1 = EV(x|\theta)$ and V_{xx} have the same structure as an intraclass correlation matrix. In particular:

$$V_{xx} = (m_2 - m_3)\, I_{(n)} + (m_3+v_{11})\ \mathbf{1}\mathbf{1}' \tag{2.29}$$

Formula (2.16) specializes then as follows:

$$\hat{\mathrm{E}}(\theta|x) = m+[m_2+(n\text{-}1)m_3+nv_{11}]^{-1}v_1\mathbf{1}'[x\text{-}m_1\mathbf{1}] \tag{2.30}$$

We note that (2.30) is a linear function of $\bar{x}$, the sample mean, ($\bar{x} = n^{-1}\mathbf{1}'x$); thus the L.S. *approximation of θ (or of $E(\theta|x)$) by x depends on $\bar{x}$ only.* This will be further analysed in Section 3. As this dependence is linear, we conclude:

$$\hat{E}(\theta|x) = \hat{\mathrm{E}}(\theta|\bar{x}). \tag{2.31}$$

An alternative proof of (2.31) would run as follows. Since:

$$V(\bar{x}) = n^{-2}\mathbf{1}'V_{xx}\mathbf{1} = n^{-1}[m_2\text{-}m_3+n(m_3+v_{11})], \tag{2.32}$$

formula (2.30) may be rewritten as follows:

$$\hat{\mathrm{E}}(\theta|x) = \hat{\mathrm{E}}(\theta|\bar{x}) = m+ \frac{\bar{x}\text{-}m_1v_1}{V(\bar{x})} \tag{2.33}$$

where, evidently, $v_1 = V_{\theta x}$.

From (2.17) and (2.32), the average measure of accuracy, $V(\eta)$, takes the form:

$$V(\eta) = V_{\theta\theta} \text{-}[V(\bar{x})]^{-1}\ v_1v_1' \tag{2.34}$$

The rule of additive precision in (2.18) now becomes

$$[EV(\theta|x)]^{-1} \geq V_{\theta\theta}^{-1} + \frac{1}{EV(\bar{x}|\theta)} \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & & & \cdot \\ \vdots & & & & \vdots \\ 0 & \cdots & & & 0 \end{bmatrix} \tag{2.35}$$

where:

$$EV(\bar{x}|\theta) = n^{-1}[m_2+(n\text{-}1)m_3] \tag{2.36}$$

and with equality in (2.35) if and only if $E(\theta|x)$ is linear in x. Note that the uncorrelated case (i.e., $\theta_3=0$ a.s.) does not provide substantial simplifications.

3. APPLICATION TO THE ESTIMATION OF A POPULATION MEAN

We now consider a sample $x = (x_1,\dots,x_n)'$ with sample mean $\bar{x} = n^{-1}\mathbf{1}'x$. Let θ be the only parameter of interest. If $E(\theta|x)$ is to be approximated by means of $\bar{x}$ alone, use would be made of:

$$\hat{\mathrm{E}}(\theta\,|\,\bar{x}) = \alpha + \beta\bar{x} \tag{2.37}$$

where

$$\alpha = E(\theta) - \beta E(\bar{x}) \tag{2.38}$$

$$\beta = \frac{\operatorname{cov}(\theta,\bar{x})}{V(\bar{x})} \tag{2.39}$$

and

$$E\ V(\theta\,|\,\bar{x}) \leq V(\theta) - \frac{[\operatorname{cov}(\theta,\bar{x})]^2}{V(\bar{x})} \tag{2.40}$$

Clearly this approximation is of interest when θ is the population mean in a first-order exchangeable process *i.e.* $E(x_i\,|\,\theta) = \theta$ $i = 1,\ldots,n$. In such a case, $\bar{x}$ is an unbiased estimator of θ: we may therefore pool the results of Sections 2.1 and 2.2.1., namely:

$$E(\bar{x}) = E(\theta) = m \tag{2.41}$$

$$\operatorname{cov}(\theta,\bar{x}) = V(\theta) \tag{2.42}$$

$$V(\bar{x}) = E\ V(\bar{x}\,|\,\theta) + V(\theta). \tag{2.43}$$

Therefore:

$$\hat{\mathrm{E}}(\theta\,|\,\bar{x}) = a\ m + (1\text{-}a)\bar{x} \tag{2.44}$$

where

$$a = \frac{V(\theta)}{V(\bar{x})} \qquad 1\text{-}a = \frac{E\ V(\bar{x}\copyright\theta)}{V(\bar{x})} \tag{2.45}$$

and

$$E\ V(\theta\,|\,\bar{x}) \leq \{V(\theta)^{-1} + [E\ V(\bar{x}\,|\,\theta)]^{-1}\}^{-1} \tag{2.46}$$

with equality if and only if $E(\theta\,|\,\bar{x}) = \hat{E}(\theta\,|\,\bar{x})$, (a.s.).

In general we also have:

$$E\ V(\theta\,|\,x) \leq E\ V(\theta\,|\,\bar{x}) \tag{2.47}$$

with equality if and only if $E(\theta\,|\,x) = E(\theta\,|\,\bar{x})$. (a.s.).

Therefore:

$$E\ V(\theta\,|\,x) \leq \{[V(\theta)]^{-1} + [E\ V(\bar{x}\,|\,\theta)]^{-1}\}^{-1} \tag{2.48}$$

with equality if and only if $E(\theta\,|\,x) = \hat{\mathrm{E}}(\theta\,|\,\bar{x})$. (a.s.). This allows us to state Ericson's (1969) result in the following way: If $E(\theta\,|\,x) = \hat{E}(\theta\,|\,\bar{x})$ (*i.e.* $E(\theta\,|\,x)$ is a linear function of $\bar{x}$) then $E(\theta\,|\,x)$ has the form (2.44) - (2.45). We may also add that $E\ V(\theta\,|\,x)$ is *equal* to the r.h.s. of (2.48).

If the process is second-order exchangeable we get an explicit form for $E\ V(\bar{x}\,|\,\theta)$ given in (2.36). With this expression, formula (2.44) - (2.45) reproduce Goldstein's (1975, b) Theorem 1 and formula (2.48) corrects his Corollary 1 (ii) (indeed, the l.h.s. of the inequality is actually the (predictive) expectation of the posterior variance and not the posterior variance itself). Note also that in these relationships, the second-order exchangeability adds only an explicit form for $E\ V(\bar{x}\,|\,\theta)$ and insures that $\hat{\mathrm{E}}(\theta\,|\,x) = \hat{\mathrm{E}}(\theta\,|\,\bar{x})$.

If we only know that $E(\theta\,|\,x) = \hat{\mathrm{E}}(\theta\,|\,x)$ (*i.e.* $E(\theta\,|\,x)$ is a linear function of x) then $E\ V(\theta\,|\,x)$ is equal to the r.h.s. of (2.22). In this case, second-order exchangeability guarantees that $E(\theta\,|\,x) = \hat{\mathrm{E}}(\theta\,|\,\bar{x})$ and, therefore, that $E\ V(\theta\,|\,x)$ is equal to the r.h.s. of (2.48). This appears in Ericson (1970) where exchangeability is obtained in the context of finite population.

3. LEAST-SQUARES SUFFICIENCY

In formula (2.31) we have seen a situation where the L.S. approximation depends on $\bar{x}$ only. One may try to characterize (i.e. to find necessary and sufficient conditions for) situations where L.S. approximation depends on $\bar{x}$ only. More generally, we may analyze under which conditions the L.S. approximation depends on a transformation of x only. This leads to the concept of "least squares sufficiency". Since $\hat{\mathrm{E}}(\theta\,|\,x)$ is a linear function of x, one should take care of linear transformation of x only. Hence, the following definition.

Definition Let t: $\mathbf{R}^p \to \mathbf{R}^s$ be a linear transformation of x i.e. $t = Ax$ $(A{:}sxp)$. Then t is *least-squares sufficient* if and only if $\hat{\mathrm{E}}(\theta\,|\,x) = \hat{\mathrm{E}}(\theta\,|\,t(x))$ for any x *(a.s.)*.

Theorem (characterization of L.S. sufficiency)[1]

[1] Comments by A.P. Dawid are gratefully acknowledged as they pointed out an error in a previous version.

Let $t = Ax \qquad (A : s \times p, r(A) = s)$
then the following conditions are equivalent:

(i) $\hat{E}(\theta \mid x) = \hat{E}(\theta \mid t(x))$ almost surely in x,
(ii) $C(A') \supseteq C(V_{xx}^{-1} V_{x\theta})$;
(iii) $\exists\, B\ (q \times s)$ such that $V_{\theta x} V_{xx}^{-1} = BA$,

where $C(\bullet)$ indicates the linear space generated by the columns of a matrix.

Proof

Condition (ii) is clearly equivalent to condition (iii) and condition (i) is equivalent to:

(iv) $$V_{\theta x} V_{xx}^{-1} = V_{\theta x} A'(A V_{xx} A')^{-1} A.$$

Indeed, using a notation similar to that of Section 1 we have:

$$E(t) = A\ E(x) \qquad V_{\theta t} = V_{\theta x} A' \qquad V_{tt} = A V_{xx} A'$$

As $C(V_{x\theta}) \subseteq C(V_{xx})$, the equivalence between (ii) and (iv) appears clearly once it has been noticed that $A'(A V_{xx} A')^{-1} A V_{xx}$ is a diagonal projection on $C(A')$. Condition (ii) of the theorem gives the geometric motivation of condition (iii) and is indeed equivalent to $t(x_1) = t(x_2) \Rightarrow \hat{E}(\theta \mid x_1) = \hat{E}(\theta \mid x_2)$.

Definition The statistic $t = Ax$ is *minimal* L.S. sufficient if and only if $C(A') = C(V_{xx}^{-1} V_{x\theta})$.

In other words, a minimal L.S. sufficient statistic may be constructed from any basis of $C(V_{xx}^{-1} V_{x\theta})$.

As an application we now answer the question considered at the begining of this section: under what condition is $\bar{x}$ L.S. sufficient? Direct application of the theorem leads to: $\hat{E}(\theta \mid x) = \hat{E}(\theta \mid \bar{x}) \Leftrightarrow \mathbf{1}$ generates the columns of $V_{xx}^{-1} V_{x\theta}$ i.e. $\exists b \in \mathbf{R}^q$ such that $V_{\theta x} V_{xx}^{-1} = b\mathbf{1}'$. Section 2.2 has shown up one such case (with $b = [nV(\bar{x})]^{-1} v_1$ - see formulae (2.30) and (2.32)).

Appendix: Derivation of (2.18).

Given (2.17), the inverse of $V(\eta)$ may be written as:

$$[V(\eta)]^{-1} = V_{\theta\theta}^{-1}\{I + [1 - [v_{11} + (\mathbf{1}' V_1^{-1} \mathbf{1})^{-1}]^{-1} v_1' V_{\theta\theta}^{-1} v_1]^{-1}$$

$$\cdot [v_{11} + (\mathbf{1}' V_1^{-1} \mathbf{1})^{-1}]^{-1} v_1 v_1' V_{\theta\theta}^{-1}\}.$$

Remember that v_1 is the first column of $V_{\theta\theta}$; this implies:

$$V_{\theta\theta}^{-1} v_1 = \begin{bmatrix} 1 \\ 0 \\ \bullet \\ \bullet \\ \bullet \\ 0 \end{bmatrix}$$

Therefore:

$$v_1' V_{\theta\theta}^{-1} v_1 = v_{11}$$

$$V_{\theta\theta}^{-1} v_1 v_1' V_{\theta\theta}^{-1} = \begin{bmatrix} 1 & 0 & \bullet & \bullet & \bullet & 0 \\ 0 & & & & & \bullet \\ \bullet & & & & & \bullet \\ \bullet & & & & & \bullet \\ 0 & \bullet & \bullet & \bullet & \bullet & 0 \end{bmatrix}$$

from which (2.18) is easily obtained.

ACKNOWLEDGEMENTS

Due to technical problems, the written version circulated at the International meeting on Bayesian statistics was a first draft (of August 1978). In the meantime, useful comments from D.R. Cox, J.B. Kadane and W.J. Rey were received and further work was going on. Consequently, the presentation of the first section has been completely rewritten thanks to the previous comments and those of the official discussants, M. Goldstein and P. Brown.

REFERENCES

BAILEY, A.L. (1950), Credibility Procedures, Laplace's Generalization of Bayes Rule, and the Combination of Collateral Knowledge with Observed Data. *Proceedings of the Casualty Actuarial Society*, **37**, 7-23.

BOUCHAT, A., (1977), *Théorie de la crédibilité: un point de vue non actuariel.* Mémoire du "Diplôme Spécial en Statistique". Université Catholique de Louvain.

BÜHLMANN, H., (1970), *Mathematical Methods in Risk Theory.* Berlin: Springer-Verlag.

— (1971), Credibility Procedures. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 515-525.

DE VIJLDER, FL., (1975), *Introduction aux théories actuarielles de crédibilité.* Office des assureurs de Belgique.

DIACONIS, P. and YLVISAKER, D., (1979), Conjugate Priors for Exponential Families. *Ann. Statist.* 7, 269-281.

DICKEY, J.M., (1969), Smoothing by Cheating. *Ann. Math. Stat.* **40**, 1477-1482.

DOOB, J.L., (1953), *Stochastic Processes.* New York: Wiley.

ERICSON, W.A., (1969), A note on the Posterior Mean a Population Mean. *J. Roy. Statist. Soc. B,* **31**, 332-334.

— (1970), On the Posterior Mean and Variance of a Population Mean. *J. Amer. Statist. Assoc.* **65**, 649-652.

FINUCAN, H.M., (1971), Posterior precision for Non-Normal Distribution. *J. Roy. Statist. Soc. B,* **33**, 95-97.

GOEL, P.K. (1979), Linear Posterior Expectation in a Scale Parameter Family and the Gamma Distribution. *Technical Report* **163**. Department of Statistics, Carnegie-Mellon University.

GOEL, P., and DeGROOT, H.M., (1979), Only Normal Distribution have Linear Posterior Expectations in Linear Regression. *Technical Report* **157**, Department of Statistics, Carnegie-Mellon University.

GOLDSTEIN, M., (1975a), Approximate Bayes Solutions to Some Non-Parametric Problems. *Ann. Statist.* **3**, 512-517.

— (1975b), A Note on Some Bayesian Non-Parametric Estimates. *Ann. Statist.* **3**, 736-740.

— (1976), Bayesian Analysis of Regression Problems. *Biometrika,* **63**, 51-58.

HARTIGAN, J.A., (1969), Linear Bayesian Methods. *J. Roy. Statist. Soc. B,* **31**, 446-454.

HEWITT, E., and SAVAGE, L., (1955), Symmetric Measures on Cartesian Products. *Trans. Amer. Math. Soc.* **80**, 470-501.

JEWELL, S.W., (1974a), The Credible Distribution. *ASTIN Bulletin* **7**, 237-269.

— (1974b), Credible Means are Exact Bayesian for Exponential Families. *ASTIN Bulletin* **8**, 77-90.

— (1974c), Exact Multidimensional Credibility. *Mitt. der Verein. Schweiz. Versich-Math.* **74**; 193-314.

KAGAN, A., LINNIK, Y.V. and RAO, C.R., (1973), *Characterizations Problems is Mathematical Statistics.* New York: Wiley.

KAHN, P.M., (1975), *Credibility Theory and Application.* New York: Academic Press.

LUKACS, E. and LAHA, R.G. (1964), *Applications of Characteristic Functions.* London: Griffin.

MAYERSON, A.L., (1964), A Bayesian View of Credibility. *Proceedings of the Casualty Actuarial Society,* **51**, 85-104.

RAIFFA, H. and SCHLAIFER, R., (1961), *Applied Statistical Decision Theory.* Harvard: University Press.

STONE, M., (1963), Robustness of Non-Ideal Decision Procedures. *J. Amer. Statist. Assoc.* **58**, 480-486.

Approximate Bayesian Methods

D.V. LINDLEY
University College London

SUMMARY

This paper develops asymptotic expansions for the ratios of integrals that occur in Bayesian analysis: for example, the posterior mean. The first term omitted is $0(n^{-2})$ and it is shown how the term $0(n^{-1})$ can be of importance.

Keywords: ASYMPTOTIC EXPANSIONS; STEEPEST DESCENT; BAYESIAN METHODS; POSTERIOR MOMENTS; ONE-WAY ANALYSIS OF VARIANCE.

1. GENERAL DEVELOPMENT

In this paper we discuss the approximate evaluation of the ratio of integrals of the form

$$\int w(\theta)e^{L(\theta)}d\theta / \int v(\theta)e^{L(\theta)}d\theta. \qquad (1)$$

Here $\theta = (\theta_1, \theta_2,\ldots,\theta_m)$ is a parameter and

$$L(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

is the logarithm of the likelihood for n observations $x_1, x_2,\ldots,x_n$, forming a random sample from a density $p(\cdot|\theta)$. The functions $w(\cdot)$ and $v(\cdot)$ are arbitrary. A simple example is where $w(\theta) = \theta_s v(\theta)$ and $v(\cdot)$ is a prior distribution for θ, when (1) is the posterior mean of θ_s. Notice that the notation $L(\theta)$ suppresses the dependence on $x_1, x_2,\ldots,x_n$. This is convenient because, in a Bayesian analysis, the x's, as observed data, are fixed and variation with respect to them is of no interest.

We shall be concerned with the asymptotic behaviour as $n \to \infty$ under regularity conditions, which will not be spelt out, in which $L(\theta)$ concentrates around the unique maximum likelihood value $\hat{\theta} = \hat{\theta}(x_1, x_2,\dots,x_n)$, obtaining an asymptotic series in inverse powers of n as far as the term of order n^{-1}. Integrals of the form occurring in the numerator and denominator of (1) were considered by Lindley (1961) for univariate θ, $(m=1)$. He obtained asymptotic expansions as far as the term of order n^{-1} We here show that the asymptotic results for *ratios* of integrals are simpler than those for the separate integrals; and we illustrate the use of the expansions in several situations.

In the multivariate case the notation requires care. The basic idea is to expand the functions involved about $\hat{\theta}$ so obtaining terms involving $(\theta_i\text{-}\hat{\theta}_i)$, $(i=1, 2,\dots,m)$. We write this deviation simply as θ_i, effectively using $\hat{\theta}_i$ as the origin. Many partial derivatives occur and we write, for example, $\partial^3 L/\partial\theta_i\partial\theta_j\partial\theta_k$ as L_{ijk}. Hence each suffix denotes differentiation once with respect to the variable having that suffix. Thus L_{222} is the third derivative with respect to θ_2. All these are evaluated at $\hat{\theta}$. Notice that the order of the suffixes is irrelevant. Similar notations are used for v and w. With these conventions, the Taylor series expansion for L, say, about $\hat{\theta}$ may be written

$$L(\theta) = L + \Sigma L_i\theta_i + \tfrac{1}{2!}\Sigma L_{ij}\theta_i\theta_j + \tfrac{1}{3!}\Sigma L_{ijk}\theta_i\theta_j\theta_k + \dots$$

where all summations run over all suffixes from 1 to m, the dimensionality of θ. We begin by considering the numerator of (1) deriving the multivariate extension of the univariate results of Lindley (1961). It is important in collecting terms of like order together, to remember that L, and all of its derivatives, are $0(n)$, whereas θ_i, for all i, is $0(n^{-1/2})$. On expansion to $0(n^{-1})$ we have

$$\begin{aligned}
&\int w(\theta)e^{L(\theta)}d\theta \\
&\quad= \int \left[w + \Sigma w_i\theta_i + \tfrac{1}{2!}\Sigma w_{ij}\theta_i\theta_j + \dots\right] \exp\left[L + \Sigma L_i\theta_i + \tfrac{1}{2!}\Sigma L_{ij}\theta_i\theta_i + \right. \\
&\qquad\qquad \left.\tfrac{1}{3!}\Sigma L_{ijk}\theta_i\theta_j\theta_k + \tfrac{1}{4!}\Sigma L_{ijkl}\theta_i\theta_j\theta_k\theta_l + \dots\right]d\theta \\
&\quad= we^L\int\left[1 + \Sigma W_i\theta_i + \tfrac{1}{2}\Sigma W_{ij}\theta_i\theta_j + \dots\right] \exp\left[\tfrac{1}{2}\Sigma L_{ij}\theta_i\theta_j\right] \\
&\quad\times \left[1 + \tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k + \tfrac{1}{24}\Sigma L_{ijkl}\theta_i\theta_j\theta_k\theta_l + \tfrac{1}{2}\left\{\tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k\right\}^2 + \dots\right]d\theta.
\end{aligned}$$

Here $W_i = w_i/w$, etc., $L_i = 0$, since the expansion is about the maximum likelihood value, and all functions are evaluated at $\hat{\theta}$. It is assumed that $w = w(\hat{\theta})$ does not vanish: the case where it is zero will be discussed below. Collecting terms of like order together, the integral is easily seen to be

$$\begin{aligned}
we^L\int e^{\Sigma L_{ij}\theta_i\theta_j/2}&\left[1 + \Sigma W_i\theta_i + \tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k + \tfrac{1}{2}\Sigma W_{ij}\theta_i\theta_j\right. \\
&\left.+ (\Sigma W_i\theta_i)\tfrac{1}{6}\Sigma L_{ijk}\theta_i\theta_j\theta_k + R\right]d\theta.
\end{aligned}$$

The orders of the terms in square brackets are respectively 1, $n^{-1/2}$, $n^{-1/2}$, n^{-1}, n^{-1} and n^{-1}, with the final term R not involving W or its derivatives. In subsequent calculations R will disappear, so we have not spelt it out.

The integrations all involve the moments of the multivariate normal distribution with density proportional to $\exp((1/2)\Sigma L_{ij}\theta_i\theta_j)$. The precision matrix has elements $-L_{ij}$. The elements of the matrix inverse to this are written σ_{ij}, forming a matrix Σ. It is well-known that for this distribution, $E(\theta_i) = 0$, $E(\theta_i\theta_j) = \sigma_{ij}$ and $E(\theta_i\theta_j\theta_k) = 0$. It is not perhaps so well-known that $E(\theta_i\theta_j\theta_k\theta_l) = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$: see, for example, Anderson (1958: equation (26) of §2.6). The result of the integration is that

$$\begin{aligned}
&\int w(\theta)e^{L(\theta)}d\theta \sim we^L(2\pi)^{m/2}|\Sigma|^{1/2} \text{ x} \\
&[1 + \tfrac{1}{2}\Sigma W_{ij}\sigma_{ij} + \tfrac{1}{6}\Sigma L_{ijk}W_l(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}) + R^*],
\end{aligned} \tag{2}$$

where R^* arises from R: the terms in square brackets being of order n^{-1} apart from the first. The second summation can be simplified since all three terms in it are equal. To see this, remember L_{ijk} is unaffected by permutation of its suffixes, so that permuting j and k in the first term gives $\Sigma L_{ikj}W_l\sigma_{ij}\sigma_{kl}$, and then interchanging the roles of j and k makes this equal to $\Sigma L_{ijk}W_l\sigma_{ik}\sigma_{jl}$, the second term: the third follows similarly.

This result is of interest in its own right but is complicated if the term R^* is spelt out. However, if we pass to a ratio (1), of such integrals with the same likelihood, the terms outside the square brackets in (2) cancel except for $w = w(\hat{\theta})$ and $v = v(\hat{\theta})$: and on expanding the ratio of the two terms in square brackets to order n^{-1}, R^*, which does not involve w, cancels with the *same* term R^* in the denominator, so that finally we have

$$\begin{aligned}
&\int w(\theta)e^{L(\theta)}d\theta/\int v(\theta)e^{L(\theta)}d\theta \sim \\
&\frac{w}{v}\left[1 + \tfrac{1}{2}\Sigma(W_{ij}\text{-}V_{ij})\sigma_{ij} + \tfrac{1}{2}\Sigma L_{ijk}(W_l\text{-}V_l)\sigma_{ij}\sigma_{kl} + \dots\right].
\end{aligned}$$

(It has been assumed that $v \neq 0$.)

In the applications we have in mind, $v(\theta) = \pi(\theta)$, the prior distribution for θ , so that the denominator is the normalizing constant in Bayes theorem; and $w(\theta) = u(\theta)\pi(\theta)$, so that the ratio is $E[u(\theta)|x_1, x_2,\dots,x_n]$. Simple calculation then shows that $W_{ij}\text{-}V_{ij} = u_{ij}/u + (u_i\pi_j + u_j\pi_i)/u\pi$ and $W_i\text{-}V_i = u_i/u$. If we write $\varrho(\theta) = \log\pi(\theta)$, a little more calculation finally gives

$$\int u(\theta)e^{L(\theta)+\rho(\theta)}d\theta \Big/ \int e^{L(\theta)+\rho(\theta)}d\theta \sim$$

$$u+\tfrac{1}{2}\,\Sigma(u_{ij}+2u_i\varrho_j)\sigma_{ij} + \tfrac{1}{2}\Sigma L_{ijk}u_l\sigma_{ij}\sigma_{kl} \tag{3}$$

to order n^{-1}. This is our basic result. The first term is 0(1) : the next are all $0(n^{-1})$ and will be referred to as *correction* terms. Notice that, because of the vanishing of all moments of odd orders for a multivariate normal distribution, the first term neglected is $0(n^{-2})$, not $0(n^{-3/2})$. Remember that on the right-hand side of (3) all functions are evaluated at the maximum likelihood value of θ, and that summation is over all suffixes and from 1 to m. One feature of immediate interest in (3) is that it does not involve the second derivatives of the prior, but that those of u do occur. Secondly, the prior is absent from the last correction term incorporating the third derivatives of the log-likelihood.

An alternative form is available for the final term in (3). Since the matrix of elements σ_{ij} is inverse to that of elements $-L_{ij}$, we have $\sum_k L_{ik}\sigma_{kl} = -\delta_{il}$. On differentiating with respect to θ_j, we obtain $\sum_k L_{ijk}\sigma_{kl} + \sum_k L_{ik}(\sigma_{kl})_j = 0$. Hence

$$\begin{aligned}\Sigma L_{ijk}u_l\sigma_{ij}\sigma_{kl} &= -\Sigma u_l\sigma_{ij}L_{ik}(\sigma_{kl})_j\\ &= \Sigma u_l\delta_{ik}(\sigma_{kl})_j, \text{ on summing over } i,\\ &= \sum_{k,l} u_l(\sigma_{kl})_k, \text{ on summing over } j. \end{aligned}\tag{4}$$

Although it appears simpler, we have found this form less convenient than that in (3) because it uses the algebraic inversion of $-L_{ij}$, in order to find $(\sigma_{kl})_k$, whereas the other only requires the numerical inversion in any application.

Another form of (3) may be obtained by writing $\Lambda(\theta) = L(\theta) + \varrho(\theta)$ which, apart from an additive constant, is the logarithm of the posterior distribution of θ, given $x_1, x_2,\ldots,x_n$. Then, instead of expanding about the maximum likelihood value, $\Lambda(\theta)$ may be expanded about its maximum, the posterior mode. Consideration of each of the individual steps in the argument that led to (3) shows that they apply when Λ replaces L. Effectively in (1), v becomes 1 and w, u. Hence

$$\int u(\theta)e^{\Lambda(\theta)}d\theta \Big/ \int e^{\Lambda(\theta)}d\theta \sim$$

$$u+\tfrac{1}{2}\Sigma\ u_{ij}\tau_{ij} + \tfrac{1}{2}\Sigma\Lambda_{ijk}u_l\tau_{ij}\tau_{kl}. \tag{5}$$

Here $\tau_{ij} = -\Lambda^{ij}$ and all quantities are evaluated at the posterior mode, $\tilde{\theta}$, instead of the maximum likelihood value, $\hat{\theta}$. An alternative form is available using a result parallel to (4). (5) is simpler than (3), but the latter has the advantage of explicitly displaying the separate roles of u and π.

An important special case is where $u(\theta) = \theta_s$, $1 \leq s \leq m$, so that the ratio of integrals is the posterior mean of θ_s, $\bar{\theta}_s$ say. Since $u_s = 1$, $u_t = 0$ for $t \neq s$ and $u_{ij} = 0$, (5) immediately shows that the difference between the posterior mean and mode for θ_s is

$$\bar{\theta}_s - \tilde{\theta}_s \sim \tfrac{1}{2}\sum_{i,j,k} \Lambda_{ijk}\tau_{ij}\tau_{ks}. \tag{6}$$

A similar result for the maximum likelihood values is, from (3),

$$\bar{\theta}_s - \hat{\theta}_s \sim \Sigma\varrho_i\sigma_{is} + \tfrac{1}{2}\sum_{i,j,k} L_{ijk}\sigma_{ij}\sigma_{ks}. \tag{7}$$

Similar calculations using $u(\theta) = \theta_s\theta_t$ give results which, when combined with (6), show that the posterior dispersion matrix for θ has elements τ_{ij} to $0(n^{-1})$, so requiring no correction from the corresponding modal values. Equivalent use of (3) shows that τ_{ij} may be replaced by σ_{ij} to the same order. Thus there is an order n^{-1} correction to the mean but not to the dispersion. An alternative way of obtaining this result is to use $u(\theta) = (\theta-\hat{\theta}_s)(\theta-\hat{\theta}_t)$, but this, and its first derivatives, vanish at $\hat{\theta}$, so that our expressions are no longer valid. The modifications necessary in this case are a little tedious, though straightforward in principle, and we therefore do not provide a general treatment but discuss special cases below: from these, the reader will be able to see how a general discussion would proceed.

The results simplify if the parameters are locally orthogonal: that is, if $L_{ij} = 0$, and hence $\sigma_{ij} = 0$, for all $i \neq j$. For example, the right-hand side of (3) reduces to

$$u + \tfrac{1}{2}\Sigma(u_{ii}+ 2u_i\ \varrho_i)\sigma_{ii} + \tfrac{1}{2}\Sigma L_{iik}u_k\sigma_{ii}\sigma_{kk}\ ,$$

and (7), for the mean, is simply

$$\bar{\theta}_s - \hat{\theta}_s \sim \varrho_s\sigma_{ss} + \tfrac{1}{2}\Sigma L_{iis}\sigma_{ii}\sigma_{ss}.$$

Local orthogonality can always be obtained by a locally orthogonal transformation of the parameter space at $\hat{\theta}$, or $\tilde{\theta}$.

Parameters are usually said to be orthogonal if $EL_{ij}(\theta) = 0$ for all $i \neq j$ and all θ; the expectation being over $x_1, x_2,\ldots,x_n$ (Jeffreys (1961)). Since L and its derivatives are sums of n terms, and hence of order n, they will, by the central limit theorem, differ from their expectations by a term of order $n^{-1/2}$. Hence replacement of L_{ij}, or L_{ijk}, by expectations will not, as many writers have noticed, affect the order of the correction terms, but it will affect the order of the terms discarded. As pointed out above, at the moment these are $0(n^{-2})$: if expectations are used they will rise to $0(n^{-3/2})$. Consequently the replacements should be used with care. Actually they violate the likelihood principle and are hence incoherent. In any case, as we try to show by example below, they are not needed in the numerical analysis of data. If they are used and the parameters are orthogonal, then further reductions occur: (3) reducing to

$$E(u) \sim u + \tfrac{1}{2}\Sigma(u_{ii}+2u_i\varrho_i)\sigma_{ii} + \tfrac{1}{2}\Sigma L_{iii}u_i\sigma_{ii}^2$$

and (7) to

$$\bar{\theta}_s - \hat{\theta}_s = \varrho_s\sigma_{ss} + \tfrac{1}{2}L_{sss}\sigma_{ss}^2 .$$

These reductions arise because the vanishing of the mixed second derivatives for all θ implies zero values for the mixed third derivatives.

An obvious advantage of some form of orthogonality is the diagonal form of the matrix of elements $-L_{ij}$ and the consequent ease of its inversion to give σ_{ij} : $\sigma_{ii} = -L_{ii}^{-1}$ and $\sigma_{ij} = 0$ for $i \neq j$.

But an additional advantage is the reduction in the numbers of third derivatives that have to be considered. These are $m(m+1)(m+2)/6$ if all distinct ones are needed; m^2 with local orthogonality; and m with full orthogonality. Full orthogonality cannot usually be achieved for $m>3$.

2. UNIVARIATE APPLICATIONS

In this section the case is considered of a single parameter, written θ, hence $m = 1$. The notation L_{ijk} etc., for the derivatives is cumbersome, all suffixes necessarily being 1, and we revert to the more usual form in which L_3, for example, denotes the third derivative; previously L_{111}. The basic result (3) is that

$$E(u\,|\,x_1, x_2,\ldots,x_n) \sim u + \tfrac{1}{2}(u_2+2u_1\varrho_1)\sigma^2 + \tfrac{1}{2}L_3u_1\sigma^4 \tag{8}$$

whereas in posterior mode form (5)

$$E(u\,|\,x_1, x_2,\ldots,x_n) \sim u + \tfrac{1}{2}u_2\tau^2 + \tfrac{1}{2}\Lambda_3u_1\tau^4 \tag{9}$$

The results for $u(\theta) = \theta$, giving the posterior mean $\bar{\theta}$, are

$$\bar{\theta} - \hat{\theta} = \varrho_1\sigma^2 + \tfrac{1}{2}L_3\sigma^4 \tag{10}$$

and

$$\bar{\theta} - \tilde{\theta} = \tfrac{1}{2}\Lambda_3\tau^4 \tag{11}$$

It is clear from these formulas that there would be some advantage in arranging for L_3, or Λ_3, to be zero. This can be done in the case of the exponential family with a single sufficient statistic. In the canonical form, the density exp $[-x\theta - g(\theta) - h(x)]$ gives a log-likelihood $L(\theta) = -X\theta - ng(\theta)$ with $X = \Sigma x_i$ the sufficient statistic, and $L_i = -ng_i$ for $i > 1$, irrespective of the sample values. Suppose the parameterization is altered from θ to ϕ where $d\phi/d\theta = L_2^{1/3}$. Then $d\theta/d\phi = L_2^{-1/3}$ and $d^2\theta/d\phi^2 = -\frac{1}{3}L_3/L_2^{5/3}$. Consequently

$$\frac{d^3L}{d\phi^3} = L_3\left(\frac{d\theta}{d\phi}\right)^3 + 3L_2\frac{d\theta}{d\phi}\frac{d^2\theta}{d\phi^2}\,;$$

since $L_1 = 0$, vanishes. Hence a change from the canonical parameter θ to ϕ, where $d\phi/d\theta = L_2^{1/3}$, or $\phi = \int L_2^{1/3}(\theta)d\theta$ will make the final correction terms in (8) and (10) vanish. If the conjugate family is used for the prior to the exponential family, the same arguments will apply to Λ and, from (11), the posterior mean and mode will be the same to order n^{-1}.

As an example consider the gamma distribution with $p(x|\theta) \sim \theta^re^{-\theta x}$; $g(\theta) = -r \log \theta$, so that $L_2 = ng_2 = nr\theta^{-2}$. Then $d\phi/d\theta = \theta^{-2/3}$, the constant being irrelevant, and hence $\phi = \theta^{1/3}$. With this parametric form, $L(\phi) = -X\phi^3 + 3nr \log \phi$ and $d^3L/d\phi^3 = 0$. This is the Wilson-Hilferty transformation, though applied to the parameter rather than the data.

It is a curious feature of the exponential family that in canonical form the derivatives of the log-likelihood above the first do not involve the data. An important effect of this is that the sampling theorist's violation of the likelihood principle in taking expectations over the sample space does no damage to the principle when applied to these higher derivatives: in particular, the large-sample variance, $\sigma^2 = -L_2^{-1}$, is unaffected. In general the derivatives will be data dependent and a transformation that makes L_3 zero is not available. An argument similar to that used above shows that a change to $\phi = \int \{EL(\theta)\}^{1/3}\, d\theta$ will make $EL_3 = 0$. As explained above, a change from L_i to EL_i will change the order of the neglected terms.

Transformations associated with L_3 are sometimes used to control skewness. It is therefore of interest to examine the third moment of θ. To do this we need the case $u(\theta) = (\theta - \hat{\theta})^3$ in the univariate form of (3). But $u = u(\hat{\theta})$ vanishes and our results do not apply. We therefore develop an expansion analogous to (2) valid when $w = 0$, confining ourselves to the univariate case. Multivariate extensions follow straightforwardly. Suppose that the first non-vanishing derivative of w at $\hat{\theta}$ is the s^{th}, $s>0$. The derivatives will be written w_s etc. Then as in the derivation of (2)

$$\int w(\theta)e^{L(\theta)}d\theta = \int\left[\frac{1}{s!}w_s\theta^s + \frac{1}{(s+1)!}w_{s+1}\theta^{s+1} + \ldots\; e^{L(\theta)}\right]d\theta$$

$$= \frac{w_s e^L}{s!}\left[\int \theta^s + \frac{W_{s+1}}{s+1}\theta^{s+1} + \dots\right] e^{(1/2)L_2\theta^2}\left[1 + \frac{1}{6}L_3\theta^3 + 0(n^{-1})\right] d\theta$$

There are two cases according as s is odd or even. In the even case the leading term is $w_s e^L\sqrt{2\pi}\sigma E(\theta^s)/s!$. In the odd case, two terms need consideration and we have

$$\left\} w_s e^L \sqrt{2}\pi\ \sigma/s! \right\{ \left\{ \frac{W_{s+1}}{s+1} E(\theta^{s+1}) + \frac{1}{6} L_3 E(\theta^{s+3}) \right\{$$

We next need to combine the results for the numerator, for w, with those for the denominator, for v. In applications $v = e^\varrho$ is the prior. We shall suppose that this nowhere vanishes, in line with the principle that a Bayesian should never assign zero probability to any value, because to do so would commit him to zero irrespective of any data. This being so, the dominant term in the denominator is $ve^L\sqrt{2\pi}\ \sigma$ giving

$$\frac{\int w(\theta)e^{L(\theta)}d\theta}{\int v(\theta)e^{L(\theta)}d\theta} \sim \begin{cases} w_s E(\theta^s)/s!v & \text{s even} \\ \{w_{s+1}E(\theta^{s+1})/(s+1) + w_s L_3 E(\theta^{s+3})/6\}/s!v, & \text{s odd} \end{cases} \tag{12}$$

of order $n^{-s/2}$ for s even, and $n^{-(s+1)/2}$ for s odd.

To obtain the posterior moments we write $w(\theta) = (\theta\text{-}\hat{\theta})^s e^{\varrho(\theta)}$ and $v(\theta) = e^{\varrho(\theta)}$. For $s = 2$, we immediately obtain σ^2, a result discussed in the general development. The third moment is a little more complicated. The first non-vanishing derivative is $w_3 = 3!e^\varrho$ and $w_4 = 4!e^\varrho\varrho_1$. Hence

$$E(\theta\text{-}\hat{\theta})^3 \sim E(\theta^4)\varrho_1 + \tfrac{1}{6}L_3E(\theta^6) = 3\sigma^4\varrho_1 + \tfrac{5}{2}\sigma^6 L_3,$$

of order n^{-2}. The fourth moment is easily seen to $3\sigma^4$. To obtain the moments about the mean write

$$\begin{aligned} E(\theta\text{-}\bar{\theta})^3 &= E(\theta\text{-}\hat{\theta} + \hat{\theta}\text{-}\bar{\theta})^3 \\ &= E(\theta\text{-}\hat{\theta})^3 + 3E(\theta\text{-}\hat{\theta})^2(\hat{\theta}\text{-}\bar{\theta}) + 2(\bar{\theta}\text{-}\hat{\theta})^3 \\ &= 3\sigma^4\varrho_1 + \tfrac{5}{2}\sigma^6 L_3 + 3\sigma^2\{-\varrho_1\sigma^2 - \tfrac{1}{2}L_3\sigma^4\} + 0(n^{-3}) \\ &= L_3\sigma^6 + 0(n^{-3}), \end{aligned} \tag{13}$$

also of order n^{-2}. Similarly $E(\theta\text{-}\bar{\theta})^4 = 3\sigma^4 + 0(n^{-3})$.

It is interesting to see that neither of these involve the prior distribution and that the fourth moment is that predicted by assuming a normal distribution for θ. Skewness would seem to be a more important feature of posterior distributions than kurtosis.

We now consider some examples, excluding the exponential family which, as we have seen, is somewhat unusual. The first is a sample from a t-distribution of unknown location, but known spread and degrees of freedom: the sample size is $n = 7$, and the degrees of freedom are 5. The log-density for x is therefore $C - 3\log\{1 + (x\text{-}\theta)^2/5\}$. With true value $\theta = 0$ the sample is:

$$-1.0\ ;\ -0.3\ ;\ -0.1\ ;\ +0.4\ ;\ +0.9\ ;\ +1.6\ ;\ +3.0 \tag{14}$$

The upper 1% point of t_5 is 3.36, so that the last value is unusual and almost deserves the title of an outlier: it would certainly be an outlier for the corresponding normal distribution with $\nu = \infty$. Table 1 gives the value of the log-likelihood and its first three differences around the maximum value. Interpolation gives $\hat{\theta} = 0.4954$, and L_2 at this value is -4.923 from the second differences. Hence $\sigma^2 = 0.2031$ and $\sigma = 0.451$. Simple calculation for the t-distribution shows that $E(L_2) = -n(\nu+1)/(\nu+3)$, giving here an average value of σ^2 of 0.190, slightly less than the sample value obtained here, so that the sample is a little less informative than an average one. Assuming $\varrho_1 = 0$ corresponding to a *flat* prior at $\hat{\theta}$, the correction for $\hat{\theta}$, equation (10), is $\tfrac{1}{2}L_3\sigma^4$. With $L_3 = 0.724$, by interpolation in the third differences, the correction to $\hat{\theta}$ is 0.0149, so that $\bar{\theta} = 0.5103$. The correction is negligible in comparison with the standard deviation. Notice, however, that $\bar{\theta}$ is very different from the arithmetic mean of the sample at 0.643, which is unduly swayed by the outlier. The correction for the prior need not be negligible. Suppose that $\pi(\theta)$ is such that θ/K is t_ν: that is, centred at the true value of $\theta = 0$ but with variance $K^2\nu/(\nu\text{-}2)$ for $\nu > 2$. It is easy to establish that $\varrho_1 = -\hat{\theta}(\nu+1)/(K\nu+\hat{\theta}^2)$. For example with $K = 1$ and $\nu = 5$, roughly making the prior equivalent to an extra value at 0, the correction term $\varrho_1\sigma^2 = -0.115$, giving $\bar{\theta} = 0.395$. Increasing K to 2, making one initially less sure about θ, gives a correction of -0.0589 and $\bar{\theta} = 0.451$.

The general form of the correction to $\bar{\theta}$ due to the prior, $\varrho_1\sigma^2$, is best appreciated by the following heuristic argument. In a quadratic approximation to the logarithm, ϱ, of the prior, it can be written $-(\theta\text{-}\theta_0)^2/2\sigma_0^2$ where θ_0 is the prior mean (or mode) and σ_0^2 the prior variance. Its derivative at $\hat{\theta}$ is $-(\hat{\theta}\text{-}\theta_0)/\sigma_0^2$. Hence, ignoring the other correction term $\tfrac{1}{2}L_3\sigma^4$.

$$\bar{\theta} \sim \hat{\theta} + \sigma^2(\theta_0\text{-}\hat{\theta})/\sigma_0^2 \sim \{\hat{\theta}/\sigma^2 + \theta_0/\sigma_0^2\}\{\sigma^{-2} + \sigma_0^{-2}\}^{-1}$$

to order n^{-1}. This is the usual weighted average of $\hat{\theta}$ and θ_0 with weights equal to their precisions.

Table 1. $L(\theta) = -3\Sigma \log \{1 + (x_i - \theta)^2/5\}$ and its differences for the sample (14).

θ	L	L_1	L_2	L_3
0.475	-4.869 031 221			
		+ 759 378		
485	8 271 843		- 492 996	
		+ 266 382		+ 705
495	8 005 461		- 492 291	
		- 225 909		+ 741
505	8 231 370		- 491 550	
		- 717 459		
515	8 948 829			

Returning to the sample (14), consider what happens when the outlier *increases* from 3.0 to 4.0. The maximum likelihood value *decreases* from 0.4954 to 0.4714, showing that less attention is paid to the extreme value. The variance σ^2 increases slightly from 0.2031 to 0.2058 and L_3 grows from 0.724 to 0.825, with the result that the correction $\frac{1}{2} L_3\sigma^4$ changes from 0.0149 to 0.0175. Hence $\bar{\theta} = 0.4889$. There is still little skewness in the posterior distribution. This is a result of the symmetry in the original density. To exhibit a substantial correction term it is necessary to take a skew density for θ, but before doing this there is one more remark that is worth making about the t-distribution. It can happen that the log-likelihood has two local maxima, in which case each will give a contribution in the asymptotic expansions.

To exhibit a skew distribution giving a larger correction term, consider a sample, again of size 7, from an F-distribution of unknown scale. We have taken a case with dégrees of freedom, $\nu_1 = 4$ and $\nu_2 = 8$, giving a density proportional to $\theta^2 x/(8+4\theta x)^6$. With true value $\theta = 1$, the sample is

$$.3 \quad .5 \quad .8 \quad 1.2 \quad 1.4 \quad 2.5 \quad 4.0 \qquad (15)$$

Table 2 gives the value of the log-likelihood and its first three differences around the maximum value. Interpolation gives $\hat{\theta} = 0.8110$, and L_3 at this value is -12.399. Hence $\sigma^2 = 0.08065$ and $\sigma = 0.2840$. The value of L_3 is 42.0, so that with a *flat* prior the correction term, $\frac{1}{2} L_3\sigma^4$, is 0.1366. The result of applying this is that $\hat{\theta}$ at 0.8110 is increased to $\bar{\theta}$ at 0.9476, and the correction is almost one half the standard deviation. The posterior distribution is skew to the right, the mean exceeding the mode. The third moment, equation (13), is 0.022 and the fourth 0.0195.

Notice that in doing numerical work with the results we have not used the differential calculus to evaluate $L_i(\theta)$ and then inserted the numerical values for θ (and $x_1, x_2, \ldots, x_n$): instead $L(\theta)$ has been evaluated for a range of values of θ and the differences used to obtain $L_i(\hat{\theta})$. This reduces substantially the amount of work, both analytic and numeric, and has the advantage of displaying the form of the log-likelihood where it is large.

One other application of the basic results, (8) and (9), that merits attention is to obtain the predictive distribution. Let y be an, as yet unobserved, value whose density, given θ, is $q(y|\theta)$. Often q will be p, the density leading to L, and y, equivalently, x_{n+1}, but the results are general. Then, given $x_1, x_2, \ldots x_n$, the density of y is given by (8) with $u(\theta) = q(y|\theta)$. The leading term is $q(y|\hat{\theta})$ and the correction allows for the uncertainty about θ. Moments for the predictive distribution are available if the moments of q are expressible as functions of θ. A related use is in empirical Bayes problems which have been treated by Deely and Lindley (1979).

Dunsmore (1976) writes the predictive distribution as $\int q(y|\theta)\, p(\theta|x_1, \ldots x_n)\, d\theta$ and uses asymptotic results for the posterior distribution to obtain approximations for the univariate case that are similar to those in the present paper. The main differences are that Dunsmore's asymptotic results use $\hat{\theta}$, not $\tilde{\theta}$; σ, not τ.

Table 2. $L(\theta) = 14 \log \theta - 6\Sigma \log (8 + 4\theta x_i)$ and its differences for the sample (15)

θ	L	L_1	L_2	L_3
0.79	-108.814 597 6			
		+ 20 320		
.80	2 565 6		- 12 868	
		+ 7 452		+ 428
.81	1 820 4		- 12 440	
		- 4 988		+ 414
.82	2 319 2		- 12 026	
		- 17 014		
.83	4 020 6			

3. BIVARIATE APPLICATIONS

With two parameters, θ_1 and θ_2, there are only 4 third derivatives and the notation L_{30} etc., in lieu of L_{111} etc., seems preferable. The correction term $\frac{1}{2} L_{ijk}u_l\sigma_{ij}\sigma_{kl}$ (equation (3)) becomes one half

$$L_{30}\{u_1\sigma_{11}^2 + u_2\sigma_{11}\sigma_{12}\} + L_{21}\{3u_1\sigma_{11}\sigma_{12} + u_2(\sigma_{11}\sigma_{22} + 2\sigma_{12}^2)\}$$
$$+ L_{12}\{u_1(\sigma_{11}\sigma_{22} + 2\sigma_{12}^2) + 3u_2\sigma_{12}\sigma_{22}\} + L_{03}\{u_1\sigma_{12}\sigma_{22} + u_2\sigma_{22}^2\}.$$

An alternative form uses $U_k = \sum_l u_l\sigma_{kl}$. The whole expression (3) is then

$$u + \tfrac{1}{2}\,\Sigma u_{ij}\sigma_{ij} + \Sigma U_j\varrho_j + \tfrac{1}{2}\,L_{30}\sigma_{11}U_1 + \tfrac{1}{2}\,L_{21}(2\sigma_{12}U_1 + \sigma_{11}U_2)$$
$$+ \tfrac{1}{2}\,L_{12}(\sigma_{22}U_1 + 2\sigma_{12}U_2) + \tfrac{1}{2}\,L_{03}\sigma_{22}U_2\,. \tag{16}$$

These expressions are complicated but well-adapted for numerical work. With L, u and ϱ evaluated, as in §2, on a grid of values of θ_1, θ_2 about $\hat{\theta}$, differences may again be used to form the derivatives, the matrix of minus the second derivatives inverted to give σ_{ij}, and then easy arithmetic gives the value of (16).

For the posterior mean of θ_1, say, we have $u(\theta) = \theta_1$ and hence $u_1 = 1$, $u_2 = 0$ and $u_{ij} = 0$ for all i, j. Hence (also from (7))

$$\tilde{\theta}_1 - \hat{\theta}_1' = \varrho_1\sigma_{11} + \varrho_2\sigma_{21} + \tfrac{1}{2}\,L_{30}\sigma_{11}^2 + \tfrac{3}{2}\,L_{21}\sigma_{11}\sigma_{12} + \tfrac{1}{2}\,L_{12}(\sigma_{11}\sigma_{22} + 2\sigma_{12}^2)$$
$$+ \tfrac{1}{2}\,L_{03}\sigma_{12}\sigma_{22}\,. \tag{17}$$

We illustrate these results for the analysis of a one-way table. This example differs from those studied in §2 in two respects. First, we operate directly with the posterior distribution rather than the likelihood. Second, the case is more interesting because the modal values (and the maximum likelihood ones) are known to be misleading, so that evaluation of means by methods that avoids tedious bivariate integrations may be of real value in the appreciation of data from such a table. There are possibilities of extensions to more elaborate analyses of variance.

The data x_{ij} ($i = 1, 2,\ldots m$; $j = 1, 2,\ldots n$) are, given $\{\mu_i\}$ and σ^2, independent with $x_{ij} \sim N(\mu_i, \sigma^2)$ that is, m groups with n observations in each group. For the prior density of the μ's, we suppose them i.i.d. $N(\mu, \tau^2)$, and independent of σ^2. This distribution can be thought of as part of the likelihood, in which case we have a Model II, rather than Model I, situation. Finally the distributions for σ^2, τ^2 and μ are supposed independent with $\nu_1\lambda_1/\sigma^2 \sim \chi^2_{\nu 1}$; $\nu_2\lambda_2/\tau^2 \sim \chi^2_{\nu 2}$ and μ uniform. The prior for σ^2 and τ^2 has not been expressed in the mathematically more convenient, conjugate form in terms of $\sigma^2 + n\tau^2$ since we believe that a prior depending on the sample size is unrealistic. Tedious calculations show that the joint posterior distribution of σ^2 and τ^2 has logarithm equal to a constant plus

$$-\tfrac{1}{2}\,(N\text{-}m + \nu_1 + 2)\log\sigma^2 - \tfrac{1}{2}\,(\nu_2 + 2)\log\tau^2 - \tfrac{1}{2}\,(m\text{-}1)\log(n\tau^2 + \sigma^2)$$
$$- nT^2/2(n\tau^2 + \sigma^2) - \nu_2\,\lambda_2/2\tau^2 - (S^2 + \nu_1\,\lambda_1)\,/2\sigma^2 \tag{18}$$

In the notation used above this is $\Lambda(\theta_1, \theta_2) = \Lambda(\sigma^2, \tau^2)$. The unexplained notation is $N = nm$, $nT^2 = n\Sigma(x_{i.} - x_{..})^2$ and $S^2 = \Sigma(x_{ij} - x_{i.})^2$, the between and within sums of squares. The modal values for σ^2 and τ^2 are easily found from (18), and these can be used to find approximate posterior means for the μ_i, which are weighted averages of $x_{i.}$ and $x_{..}$ with weights dependent on these modes. However the distribution (18) is skew and the preferred means may differ from their modes.

We illustrate using a numerical example with $\mu = 0$, $\sigma^2 = \tau^2 = 1$, having the hyperparameters, $\nu_1 = \nu_2 = 4$, $\lambda_1 = \lambda_2 = 1$, and with data $S^2 = 37.34372$, $T^2 = 4.556774$, for $m = 8$ and $n = 5$. Notice that the prior information about τ^2, with 4 degrees of freedom, is comparable with the information from the data, through T^2, having 7. The prior expectation of τ^2 is $\lambda_2\nu_2\,/(\nu_2 - 2) = 2$, and the standard deviation is infinite, but the mode is at $\lambda_2\nu_2\,/\,(\nu_2 + 2) = 2/3$. This does not seem unrealistic in some applications, though each case must be decided in the light of practical experience. Table 3 gives the value of $\Lambda(\sigma^2, \tau^2)$, equation (18), for a grid of values of σ^2 and τ^2 around $\sigma^2 = 1.1$ and $\tau^2 = 0.6$. Interpolation shows that the modal values are $\tilde{\sigma}^2 = 1.08$ and $\tilde{\tau}^2 = 0.59$. (The estimates obtained by equating the values of S^2 and T^2 to their expectations are for σ^2, 1.17 and for τ^2, 0.42.) To evaluate the correction terms, the differences are used to obtain the derivatives. Thus $\Lambda_{20} = \{(-0.2267 + 0.1213) - (-0.1213 + 0.1727)\}\,/0.01 = -15.60$. Similarly $\Lambda_{02} = -13.75$ and $\Lambda_{11} = +0.63$. The small value of this mixed, second derivative, in

Table 3. Values of $30 + \Lambda\,(\sigma^2, \tau^2)$, equation (18), for the values given in the text. All entries preceded by -0.

τ^2 σ^2	0.4	0.5	0.6	0.7	0.8
0.9	8728	5212	4522	5287	6847
1.0	5654	2319	1727	2548	4139
1.1	4894	1718	1213	2083	3701
1.2	5728	2693	2267	3181	4823
1.3	7683	4773	4417	5370	7033

comparison with the larger values of the unmixed ones, means that σ^2 and τ^2 are almost locally orthogonal and we will treat them as such in what follows. Extending to the third derivatives $\Lambda_{30} = 59.$, $\Lambda_{03} = 97.$ and Λ_{21} and Λ_{12} are virtually zero. Hence we may use the two univariate formulae for $E\ (\sigma^2)$ and $E\ (\tau^2)$ separately. For σ^2 the mode is 1.08 and the variance is $(-\Lambda_{20})^{-1} = 0.0641$, with standard deviation 0.253. The correction term is $\frac{1}{2}59 \text{ x } (.0641)^2 = 0.12$, raising σ^2 to 1.20 as the posterior mean. For τ^2 the mode is 0.59 and the variance is $(-\Lambda_{02})^{-1} = 0.0727$, with standard deviation 0.270. The correction term is $\frac{1}{2}97 \text{ x } (.0727)^2 = 0.26$ raising τ^2 to 0.85 as the posterior mean. Notice that the two correction terms are both positive, the means exceeding the modes, and that they are comparable with the standard deviations: for σ^2 the correction is about half the standard deviation, whilst for τ^2 they are about equal. Hence the term of order n^{-1} (for the correction) is comparable with that of order $n^{-1/2}$ (for the standard deviation). The claim sometimes made that terms of smaller order may be neglected in maximum likelihood (or maximum posterior) theory may not be true for some skew distributions. Notice, that because of the large, unmixed, third derivatives, the skewness in both parameters is quite large. It is interesting that the standard deviations of σ^2 and τ^2 are about equal (0.25 and 0.27 respectively) whereas one might have expected σ^2 to be better determined than τ^2.

4. DISCUSSION

The analytic results of this paper enable one to calculate the difference between the mean and mode of certain distributions as far as the dominant term of order n^{-1} in the sample size n. The difference involves the second and third derivatives of the log-likelihood at the mode and is in a form suitable for numerical calculation. Such calculations tentatively suggest that the differences are appreciable even in comparison with the standard deviations, but much more needs to be done before these claims can be substantiated.

The method used here is essentially that of steepest descents. This tool has been used by Barndorff-Nielsen and Cox (1979) to obtain sampling distributions that enable inferences to be made about one parameter, θ_m say, in the presence of nuisance parameters $\theta_1, \theta_2, \dots \theta_{m-1}$. It will be of interest to see how these sampling-theory approximations compare with the Bayesian results of this paper.

REFERENCES

ANDERSON, T.W. (1958). *An introduction to multivariate statistical analysis.* New York: Wiley

BARNDORFF-NIELSEN, O. and COX, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. Roy. Statist. Soc B*, **41**, 279-312

DEELY, J.J. and LINDLEY, D.V. (1979). Bayes empirical Bayes. *Tech. Report.* University of Canterbury.

DUNSMORE, I.R. (1976) Asymptotic prediction analysis. *Biometrika,* **63**, 627-630.

JEFFREYS, H. (1961) *Theory of probability.* Oxford: Clarendon Press.

LINDLEY, D.V. (1961) The use of prior probability distributions in statistical inference and decisions. *Proc 4th Berkeley Symp.* **1**, 453-468.

DISCUSSION

P.J. BROWN (*Imperial College, London*):

The paper by Mouchart and Simar gives an elegant exposition of some results consequent on assuming linear Bayes estimates as approximations to Bayes estimates. I have not undertaken a historical search of the relevant literature but I think the results of Hartigan (1969) might be particularly pertinent and the subsequent work of Goldstein (1975a, 1975b, 1976), deserves more direct incorporation.*

In these linear Bayes methods one only needs express the mean and variance of the joint distribution of θ and x. Further features are not required. Of course when the Bayes estimates are actually linear nothing is lost and in this context recently Diaconis and Ylvisaker * have shown how linearity is intimately connected with the exponential family and natural conjugate priors. When one gets away from such situations as is often necessary on various grounds, e.g., the need for fattailed priors to cope with discordant observations (Dawid, 1973 and Hill, 1974) then linear estimates are no longer adequate.

I would be grateful if Professor Mouchart would elaborate on his meaning of 'robustness'. Section 2.2.1 suggests that the less that is assigned the more robust the procedure. Perhaps indeed in the univariate or multivariate situation it is too much to specify all the means and variances. Exchangeability as in Section 2.2 is one way of reducing the problem but should one go further and allow the data to specify the hyperparameters? Efron and Morris (1973) for example introduce Empirical Linear Bayes estimation where, without specifying the distributional forms, linear Bayes coefficients are estimated from the data producing a rather non-linear estimate overall. These then are distributionally relaxed forms of the exciting but non-linear Stein-type estimates.

The interesting paper by Lindley seems potentially rather useful. There are two points that bother me. On Section 1, θ_i (deviation from the maximum likelihood estimate) is said to be $0(n^{-1/2})$. It would be good to have regularity conditions on prior, model and observed data justifying this, especially in view of the nature of the asymptotic expansions and subsequent integrations.

Furthermore, although it is nice notationally to suppress the data, the posterior moments under consideration are functions of both n, the sample size, and the data, so that perhaps more accurate expansions might be available if the data are also

* Incorporated in published version of Mouchart and Simar's paper.

considered. It might even be appropriate in some situations, as for example linear regression, to consider a norm depending on the data rather than just n. At any rate, here, with the usual essential asymmetry of design, n, the sample size, tells us only some of the story.

Overall we should perhaps await numerical comparison with exact results before embarking on these elegant approximations.

M. GOLDSTEIN (*University of Hull*):

Linear Bayes methods are an important recurrent theme in the Bayesian literature (as reflected by the diverse set of references in the paper by Mouchart and Simar). Although I am not quite sure in what sense the authors have simplified previously complicated results, it is useful to have a concise summary of some of the basic work in this area, and I have no particular technical points to make (essentially, I agree with the authors' presentation). Instead, I would like to raise a nontechnical point which puzzles me a little, namely in what, if any, sense can linear Bayes rules be said to be robust? This robustness is stressed at various points by the authors (and by others - I may have done so myself). However, all we are really saying is that the estimator and risk do not depend on many aspects of the prior distribution. But we might, and perhaps should, argue that if different plausible specifications of the full prior distribution give different estimates, then clearly the form of the prior is important, and this aspect of the problem cannot be ignored. Thus, our most conscientious specification of a full prior distribution should give a more meaningful answer than the linear rule, which may not approximate well to any of our plausible range of prior beliefs. I feel that the usefulness of the linear rules is that they say something precise and simple by carefully limiting the aspects of the problem allowed for consideration, but that it would be wrong to attribute any further properties to this approach without careful justification. Do the authors have any comments?

While it is often interesting to perform series expansions of akward integrals, and pick out important terms, to make the claim that the first term omitted is of $O(n^{-2})$ needs careful justification. Thus, if we are evaluating

$$\int_{-\infty}^{\infty} \omega(\theta)e^{L(\theta)}d\theta,$$

clearly this integral is not equal to the integral that $^{-\infty}$ we would obtain by replacing $\omega(\theta)$ and $L(\theta)$ by their respective series expansions. (Indeed, the latter integral may not even converge). For the suggested expansions to work, we must hope that we can find some value $\alpha(n)$ of $O(n^{-1/2})$ for which

$$\int_{|\theta| > \alpha(n)} \omega(\theta)e^{L(\theta)}\, d\theta$$

is of $O(n^{-2})$. Having done this, we may replace $\omega(\theta)$, $L(\theta)$ by their series expansions and retain only the leading terms. This will provide a valid evaluation of the integral between $\pm\,\alpha(n)$ to $O(n^{-2})$. However, even here, as the integrals of the leading terms are not evaluated between $\pm\alpha(n)$ but between $\pm\infty$, we must further check that the difference between the integrals of the leading terms between the different sets of limits is also of $O(n^{-2})$. I suspect that we may be able to do this in many useful cases; but it is not a question of regularity conditions so much as of rate of convergence. As a perhaps slightly unfair question, can we have some guidance as to when these conditions will hold?.

The easy way to explore an approximation is by trying it out for simple problems in which it is straghtforward to evaluate the integral and the approximation. A simple case which, I feel, yields some insight into the procedure is to suppose that we are drawing a sample of size n from a Bernoulli distribution with parameter θ, where the prior distribution for θ is a beta distribution with each parameter equal to a common value γ. As the approximation procedure essentially estimates the "correction" which should be applied, in large samples, to the maximum likelihood estimator θ in order to obtain, approximately, the posterior mean, a natural way to assess the approximation is to consider the ratio

$$r = \frac{\text{actual correction}}{\text{estimated correction}}$$

i.e. the ratio $(\hat{\theta}\text{-}E(\theta\,|\,\text{data}))/(\hat{\theta}\text{-}\widetilde{E}(\theta\,|\,\text{data}))$, where $\widetilde{E}$ is the suggested approximation to the posterior mean.

In this case, evaluating the required quantities gives

$$r = \frac{n}{n+2\gamma}$$

(One reason for choosing this example is that r does not depend on the observed number of successes, k, which facilitates a further comparison I shall make below).

Clearly r is of the right order, and as long as n is large compared to γ the approximation will work well. Also the correction is always in the right direction, though it always overestimates the values. However, there is a further, perhaps surprising, interpretation of r for this example, which may illuminate the relationship between the asymptotic approximation and the linear approximations discussed in the paper by Mouchart and Simar. In this problem, the posterior mean is the linear Bayes rule, i.e. the best rule of the form $a\,(k/n) + (1\text{-}a)E\theta$. The value a in this case is precisely the value r given above. Qualitatively, this gives an insight into the range of application of the two approximations. The asymptotic approximation is useful when r is near one, i.e. when a is near one and the linear Bayes rule is near θ. Thus, when n is such that the linear Bayes rule gives negligable weight to the prior mean (i.e. specification of the prior mean conveys very little conformation about the posterior mean), then it is the derivatives around θ which convey useful prior information about the posterior mean. Two further (unfair) questions. Firstly, do these qualitative insights extend to more complicated circumstances, and in particular to multiparameter problems? Secondly, should I be surprised that r is precisely equal to a (i.e. what basic property of the example I chose made it work)?.

J.M. BERNARDO (*Universidad de Valencia*):

Professor Lindley has provided us with asymptotic expansions for often encountered ratios of integrals to order $O(n^{-1})$. Nevertheless, I would like to know more about the question of when n is large enough for the approximation to be used. A formal answer surely depends on the specific problem and on the loss structure attached to the 'distance' between the true value of the ratio of the integrals and its approximation; however, maybe he can give us a feeling of the kind of situations where he expects the approximation to work.

A.P. DAWID (*The City University*):

Professor Lindley's investigation of higher-order approximations to posterior distributions comes at a time of renewed general interest in such approximations for sampling distributions, although I am hard put to recognise the relationship between Lindley's work and the methods of Barndorff-Nielsen and Cox (1979). It seems to me to be more in the spirit of the ideas on second order efficiency considered by Efron (1975). That paper used the idea of *statistical curvature*, a fascinating concept but one which is (as Lindley (1975) himself pointed out in his discussion on Efron) suspect for the Bayesian because of its dependence on the sample space. Nevertheless, I can't help feeling that a parallel, fully Bayesian, theory might be just around the corner, based on a likelihood analogue of curvature, just as the Bayesian first-order theory replaces expected Fisher information by observed information. Such a theory might be valuable for assessing the usefulness of approximations such as those of Mouchart and Simar.

Alternatively, analogues of saddle-point methods might yield accurate non-normal approximations for posterior distributions.

S. FRENCH (*Univesity of Manchester*):

Professor Lindley's paper on approximations to posterior expectations will undoubtedly lead to many fruitful applications. However, before the formulae are used, perhaps one or two cautionary remarks are appropriate.

The approximations required that certain derivatives be calculated and Professor Lindley suggests that the necessity of some rather horrendous differentiation can be avoided by recourse to finite difference approximations. Now, whilst it is generally easier to differentiate than to integrate a function *analytically*, the reverse is true of *numerical* differentiation and integration. Numerical differentiation is a very unstable operation, since it requires many small differences of function evaluations and so rounding error accumulate dramatically. See, e.g. Fröberg (1969), Fox and Mayers (1968) Blum (1972). Since these formulae require the functions to be differentiated numerically three times, these remarks are all the more appropriate.

Therefore, I would suggest that, when Professor Lindley's formulae are used, the functions should be differentiated analytically if at all possible. There are, after all, computer packages that will handle the algebraic operations of differentiation and provide the analytic form of the differential for the vast majority of the functions that arise. If analytic differentiation really is too difficult, then I suggest a visit to one's friendly neighbourhood numerical analyst. We complain enough of non-statisticians doing statistical analyses without consulting us, perhaps we should heed our own advice and consult the experts in numerical analysis.

I.J. GOOD (*Virginia Polytechnic and State University*):

Some of the mathematics in the paper resembles that used in the centroid method of integration of a positive function of several variables. Taylor's theorem in several variables is used and leads to the requirement of calculating the moments of the region of integration. See Good & Gaskins (1969,1971) and Good & Tideman (1978).

In one of Professor Lindley's expansions the term of order $1/n$ was appreciable compared with that of order $1/n^{1/2}$. This suggests that he should take the expansion at least to the next term to check its accuracy in this case.

J. GREN (*Econometric Institute, Warsaw*):

I would like to make a short comment on the paper by Professor Lindley. We know that the problem of multi-dimensional integration is the most difficult problem in Bayesian estimation of econometric models.

Up to now we have two main approaches or directions, to solve this difficult problem.

The first way is just improving the numerical methods for each separate multi-dimensional integral. They seems to be rather unpromising, even for the Cartesian product rule with Newton-Cotes Quadrature at each step.

The second way is to adopt the Monte Carlo method in order to estimate the value of each integral which appears in Bayesian estimation of econometric models. This is much more promising; see Kloek and Van Dick (1978).

Professor Lindley is now proposing a very good and operational approximation for the ratio of multi-dimensional integrals.

Since the ratio of such integrals plays a crucial role in Bayesian estimation technique, Lindley's paper opens a new, third way for obtaining practical results in Bayesian econometrics.

I would like to congratulate Professor Lindley for showing to us this new, very promising method.

A. O'HAGAN (*University of Warwick*):

Professor Lindley's expansions are extremely interesting and promise to become a standard technique, particularly in models with many parameters. For although then the expansions contain a great many terms, the saving over the vast number of function evaluations required for numerical integration will be enormous. The only lingering doubt here is whether the neglected $0(n^{-2})$ terms, whose number will also escalate rapidly, will cease to be negligible.

On a point of methodology it would seem most sensible to expand about the posterior mode $\tilde{\theta}$ than about $\hat{\theta}$. In his univariate example, with the sample from a *t* distribution, Professor Lindley uses the expansion about $\hat{\theta}$, and with the proper prior he obtains the approximate value of .395 for the mean $\bar{\theta}$. Yet if we expand about $\tilde{\theta}$, using equation (11) rather than (10), we find the new approximation .415. Which is better? Numerical integration confirms the mean to be .415 to three decimal places!

REPLY TO THE DISCUSSION

MOUCHART, M. (*Université Catholique de Louvain*) and SIMAR, L. (*Facultés Universitaires Saint-Louis, Bruxelles*):

Two types of topics seem to emerge from the discussion. The usefulness of Least Squares Approximation in Bayesian Analysis and the claim for robutsness of such procedures.

Let us first mention that the aim of the paper was not to justify the use of L.S. approximation: we only wanted to propose a simple and self-contained exposition of a host of results widespread in the literature. If justification was at stake, two types of arguments could be mentioned. One argument would be to consider the cases where the posterior expectation is (exactly) linear in x. In this connection, as pointed out by P.J. Brown, works like that of Diaconis and Ylvisaker (1979) should be mentioned. Another argument would be to consider whether a given situation is *close* to such a case. As pointed out by A.P. Dawid, the work by Efron (1975) appears to be relevant, in particular, it may help to appreciate the proper role of the coordinates in the choice of parameters and of statistics. In this line of thought, M. Goldstein seems to pay a special attention to the specification of the prior distribution. Although this is surely crucial, we like to insist that the structure of the problem is given by the *joint* distribution of (θ,x). Even if the sampling process is kept fixed, the question of whether the structure of the prior distribution will determine $E\ (\theta\,|x)$ to be more or less linear in x depends on the choice of coordinates.

P.J. Brown raised the question of whether, for example, fat-tailed prior distribution would endanger the use of L.S. approximation. Surely, fat tails in the prior distribution may lead to infinite Bayesian risk (under quadratic loss): in other words the problem may become meaningless. Remember that in a decision context only the product of the prior distribution and the utility function is relevant, thus the prior distribution and the utility function should be specified and discussed jointly. Even if the tails of the prior distribution are rather thick (the variance remaining finite), the linearity of $E\ (\theta\,|x)$ may not be affected: for instance in a multivariate student distribution, the regression functions are still linear; in any cases, $V(\eta)$ will always give an indication on the accuracy of the approximation.

Finally any discussion on *justifying* the use of L.S. approximations should involve the problem of robutsness. This is the second theme of the discussion.

First, a comment by P.J. Brown induces us to clarify a possible misinterpretation of section 2.2.1. Exchangeability was introduced as a generalization of i.i.d. processes. As such it appears as a minimal assumption that allows the reproduction and unification of earlier results; for this reason the hypothesis was decomposed into two steps. Apart from this purely formal aspect, exchangeable but not i.i.d. processes naturally appear e.g. in sampling from finite populations or when integrating out nuisance parameters in i.i.d. processes. In the latter case, indeed, $D\ (x\,|\theta_1)$, the data density marginalized on θ_2, a nuisance parameter, represents an exchangeable process and the results of section 2.2 may then be used to evaluate $\hat{E}(\theta_1|x)$ and to obtain in this way a more robust procedure.

Let us now discuss briefly what we mean by the robutsness of L.S.. approximations: this was indeed questioned by both P.J. Brown and M. Goldstein. These approximations act as *smoothing* procedures, i.e. they are less variable from (some) perturbations of a given problem (here, $D\ (x,\theta)$) than the exact solution. Apart from the search for computational simplification, a possible motivation may be the following: in case of a misspecification error it may be hoped that the approximate solution to a misspecified model might be better than the exact solution of this misspecified model. In any case this approximation is known to be free from systematic error ($E(\eta)\ =\ 0$) and some idea of the accuracy may be obtained in a simple way (by computing $V(\eta)$).

Finally we want to thank the discussants: their remarks provided help and opportunity in improving the presentation of our paper.

D.V. LINDLEY (*University College London*):

The most important point made by the discussants concerns the lack of rigour in the derivation of the results and the resulting vagueness about when the approximations are likely to be useful and accurate. These criticisms are correct and their implications most important; but I have to admit that I don't see how to meet them. It is notoriously difficult to assess the accuracy of any expansion without deriving some information about the magnitude of the terms neglected. Even to obtain the term of order n^{-2} would be a formidable undertaking since it would involve the evaluation of R * (equation (2)) plus other complicated terms. The lesser point of knowing just when the expansion is valid is easier but is beyond my limited mathematical abilities. My feeling at the moment is that understanding will be improved by investigating numerical cases and comparing the exact and approximate results. In this context, I am grateful to O'Hagan for evaluating one integral exactly, with the superb result that it agrees with the approximation to one part in 400. But one swallow does not make a summer and much more investigation is required. We have to be careful too in thinking that a term of order n^{-1} is necessarily less than one of order $n^{-1/2}$. The numerical illustration of the F-distribution in section 2 provides an example to the contrary; and other calculations that I have performed with the Weibull distribution (not reported in the paper) suggest that this can easily happen when skewness is present. The example that most interest me is that of the analysis of variance in Section 3 - and its possible extension to more complicated, higher-dimensional analyses. There it is not quite clear what is the n in the expansion in powers of $n^{-1/2}$, for there are two sample-size parameters; m, the number of groups, and n, the number of observations in each group. Presumably both have to tend to infinity, but does their relative speed of approach matter?

There is one comment on the approximation that can be made with some confidence; the expansion about the mode is typically better than that about the maximum likelihood value. This can be seen clearly in the case discussed by Goldstein. Using the latter he obtains a measure of quality of the approximation equal to $r\ =$

$n/(n+2\gamma)$. With the modal value, I find r to be $[n + 2(\gamma-1)]/(n+2\gamma)$. As he points out, r for the likelihood value is near to the desirable value of unity only when γ is small in comparison with n. The modal approximation only requires $n + 2\gamma$, a measure of the total information, likelihood plus prior, to be large. This observation is supported by evaluations of the next non-zero terms in the expansions, which is not too difficult in this case.

Similar remarks apply to the work of Dunsmore: he uses the modal expansion and his results are superior to those using a likelihood expansion. In particular, the latter can easily give rise to negative values when approximating a predictive distribution whereas this only happens for very small samples when using the modal values.

Too much should not be deduced from these calculations since we are here dealing with members of the exponential family, which is unusual in that the derivatives of the log-likelihood above the first are data-free. It is my guess that the results are likely to be most useful in the case where no sufficient statistics of low dimensionality exist and where the sample precision varies from sample to sample. This is why, to enlarge on Dawid's remark, the relationship, or lack of it, between the work of Barndorff-Nielsen and Cox and the results of this paper is of interest; they average over sample values and thereby lose sight of the fact that some samples are more informative than others.

I am grateful to Good for drawing my attention to the centroid method. The main difference between it and the device used in the paper is that the centroid uses a Taylor series expansion of a function, whereas I use one of the logarithm. As a result, where the centroid has moments of inertia, I have moments of distributions. The logarithmic expansion may be preferable here, where it is a sum of n terms, but the two methods are nicely complementary.

I have to confess that a visit to my friendly neighbourhood numerical analyst, as suggested by French, had not occurred to me since I did not see any problems in the evaluation of the differences that could not be solved by intelligent trial and error investigation of the log-posterior in the neighbourhood of the maximum. I did think of analytic differentiation on a computer, but previous experience with this was not encouraging. My personal predilection is for simple numerical analyses using simple computers where I have the feeling, perhaps erroneous, that I know what is going on. Packages and big computers terrify me. They are like some bureaucratic machine where workings and output are unintelligible; like the communication I had from the U.S. Internal Revenue Service which was most unclear as to whether I owed them money or they owed me. Only the subsequent arrival of a cheque clarified the matter. So far as computers are concerned, Schumacher is right; small is beautiful.

Dawid's suggestion of a likelihood analogue of curvature is intriguing. Efron's ideas are useful in discussing the merits of different estimators. But in the Bayesian approach there is only one estimator, the posterior distribution; or, if a decision problem is involved, the unique set of best acts. Hence no optimality considerations arise and thus there appears to be no need for curvature.

Brown is right to raise the question of dependence on the sample. As n increases, additional sample values are introduced, so that any detailed consideration of the limit as $n \to \infty$ must consider how the sample could change. Two possibilities are that we would have asymptotic convergence with probability one for each value of θ: or more weakly, with probability one - this being the overall probability incorporating $\pi(\theta)$.

I do not know the answers to Goldstein's questions. The second involves a very special situation which is perhaps only a curiosity. The first is important because it is in multiparameter problems that the ideas put forward might be most useful.

I am most grateful to all discussants for their sympathetic reception of what is an untidily, incomplete paper.

REFERENCES IN THE DISCUSSION

BLUM, E.K. (1972) *Numerical Analysis and Computation,* Reading, Mass. Addison-Wesley

DAWID, A.P. (1973) Posterior expectations for large observations. *Biometrika,* **60**, 664-666.

EFRON, B. and MORRIS, C. (1973) Stein's estimation rule and its competitors - an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117-30.

EFRON, B. (1975) Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3**, 1189-1217.

FOX, L. and MAYERS, D.F. (1968) *Computing Methods for Scientists and Engineers.* Oxford: University Press.

FRÖBERG, C.E. (1969) *Introduction to Numerical Analysis* (2nd edn) Reading, Mass: Addison--Wesley

GOLDSTEIN, M. (1975a). Approximate Bayes solutions to some non-parametric problems. *Ann. Statist.* **3**, 512-517.

— (1975b). A note on some Bayesian non-parametric problems. *Ann. Statist.* **3**, 736-740.

— (1976). Bayesian analysis of regression problems. *Biometrika* **63**, 51-58.

GOOD, I.J. and GASKINS, R.A.(1969) The centroid method of integration. *Nature* **222**, 697-698

— (1971) The centroid method of numerical integration. *Numerische Mathematik* **16**, 343-359.

GOOD, I.J. and TIDEMAN, T.N. (1978) Integration over a simplex, truncated cubes, and Eulerian numbers. *Numerische Mathematik,* **30**, 355-367

HARTIGAN, J.A. (1969). Linear Bayesian methods. *J. Roy. Statist. Soc. B,* **31**, 446-454.

HILL, B.M. (1974) On coherence, inadmissibility and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics,* (Fienberg, S.E. and Zellner, A. eds.) Amsterdam: North-Holland.

KLOEK, T. and VAN DIJK, H.K. (1978). Bayesian estimates of equation system parameters. An application of Integration by Monte-Carlo. *Econometrica,* **46**, 1-19.

LINDLEY, D.V. (1975) Comments on Efron (1975) *Ann. Statist.* **3**, 1222-1223.

Trab. Estadist. (1980)
Vol. 31 pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 247-292 Valencia, Spain: University Press

5. Regression and time series

INVITED PAPERS

BROWN, P.J. (*Imperial College, London*)
Aspects of multivariate regression

DEMPSTER, A.P., (*Harvard University*)
Bayesian inference in applied statistics

DISCUSSANTS

HARRISON, P.J., (*University of Warwick*)
ZELLNER, A., (*University of Chicago*)

REPLY TO THE DISCUSSION

Aspects of Multivariate Regression

PHILIP J. BROWN
Imperial College, London

SUMMARY

Important features of multivariate linear regression are emphasised and a selection of prior distributions discussed. Priors used by Brown and Zidek (1978) lead them to a class of 'Empirical' Bayes shrinkage estimates. The strength of shrinkage is examined with respect to an election forecasting example where observations obtain one after another.

Keywords: MULTIVARIATE RIDGE REGRESSION; MEAN SQUARE ERROR; PREDICTION; SHRINKING AND SMOOTHING; MINIMAX PROPERTIES; ELECTION FORECASTING.

1. INTRODUCTION

In numerous practical situations a set of q dependent variables or responses relate to a set of p 'independent' variables. One example is where votes to q parties in an election depend on votes at previous elections and socio-economic variables defining the voting units or constituencies. In particular, assuming the customary multivariate linear regression model for n observations on the q responses $\mathbf{Y}$, (nxq) satisfies

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

with $\mathbf{X}$ the (nxp) matrix of 'independent' variables, assumed fixed, and β a (pxq) matrix of unknown coefficients, with $\epsilon = (\epsilon^1, \ldots ,\epsilon^q)$, we make the usual assumptions on the error of normality with

$$E(\epsilon^j) = \mathbf{0},\ \mathrm{cov}(\epsilon^j,\epsilon^\ell) = \gamma_{j\ell}\mathbf{I}_n \tag{2}$$

Although the assumption is relaxed later, here $\Gamma = (\gamma_{j\ell})$ is taken to be known. The problem of interest is to estimate β, to predict the responses for a future set of m observations with given 'independent'variables $\mathbf{X}_0$ (mxp) and possibly attach some uncertainty to this prediction. This last thorny problem will not be examined in this paper. We concentrate on the first two; estimation of β and linear combinations of β. Implicitly, if not explicitly, quadratic loss is assumed. When this is explicit it will be necessary to be careful, when specifying the loss, about the weighting attached to each component of the matrix β.

It is well established (see for example Zellner, 1971, chapter 8) that under vague prior information on β, formally Lebesgue measure on R_{pq}, the Bayes estimate of β under a variety of loss functions is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3}$$

It is clear that, by Savages's precise measurement theorem, this estimate will be unperturbed by 'informative' prior knowledge for large n provided the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are 0 (n). However, typically this will require far more observations than the one parameter situation (see Hill, 1974). Incidentally, for sampling theorists, (3) is the best linear unbiased estimator (Rao, 1965 48c.2) and is the maximum likelihood estimator.

One way of arguing against (2), following Brown and Zidek (1980) and Sclove (1971), is to note that (3) may be written

$$\hat{\beta}^j = (\mathbf{X}^T\ \mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^j, \qquad j = 1,\ldots,q \tag{4}$$

where $\hat{\beta} = (\hat{\beta}^1,\ldots,\hat{\beta}^q)$, $\mathbf{Y} = (\mathbf{Y}^1,\ldots,\mathbf{Y}^q)$, so that $\mathbf{Y}^j$ $(nx1)$ and $\hat{\beta}^j$ pertain to the j^{th} of the q responses and thus the estimates (3), (4) take no account of $\Gamma = (\gamma_{j\ell})$, the between regressions covariance matrix. In the next section we discuss a selection of informative prior structures for β which result in estimates of β^j which utilise information across all q equations. These priors will have sizeable impact on the estimation of β as long as all eigenvalues of $\mathbf{X}^T\mathbf{X}$ are not large. The importance for predictions may be even greater if low information directions of $\mathbf{X}^T\mathbf{X}$ are directions for prediction (Goldstein andBrown, 1978).

In section 3, the class of ridge estimates derived from the prior of Brown and Zidek (1980) is described in some detail. Alternative ways of estimating the hyperparameters are also suggested. Sampling theory results on the ability of the ridge estimators to do better than least squares everywhere in the parameter space are summarised in section 4. Section 5 applies favoured multivariate ridge estimates to 'on the night' election forecasting data illustrating their behaviour as increasing information becomes available.

2. SOME INFORMATIVE PRIORS FOR β.

2.1. *Stage of prior construction*

A wide class of prior distributions can be obtained as a mixture of a multivariate normal distributions with respect to a few hyperparameters. A suitable choice of mixing distribution will lead to an overall prior distribution which is sufficiently uninformative in the tails to avoid the problems associated with large observations (Dawid, 1973 and Hill, 1974). Such priors, rational functions of $\|\beta\|$ for large $\|\beta\|$ are at least implicit in the derivation of Stein's estimator (see Stein, 1962 and Zellner and Vandaele, 1974). One rich example is given by Berger (1980), where if we assume the model $\mathbf{Z}$ is t-variate normal with mean θ and non-singular covariance matrix Σ (known), where θ is distributed t-variate normal with mean $\mathbf{m}$ and covariance matrix $B(\lambda) = \lambda^{-1}\mathbf{C} - \Sigma$ for $0 < \lambda < 1$, an unknown parameter with generalised prior density $\lambda^{(r-1\ -\ t/2)}$ With $\mathbf{m} = \mathbf{o}$, $\mathbf{C} = \Sigma = \mathrm{I}$ and $r = (t\text{-}2)/2$ the James-Stein (1961) estimator $\hat{\theta}$ of θ given by

$$\hat{\theta} = (1 - (t\text{-}2)/\|\mathbf{z}\|^2)\mathbf{z} \tag{5}$$

obtains to $o(\|\mathbf{z}\|^{-1})$. This general prior behaves like $k(\theta' C^{-1}\theta)^{-r}$ for large $\|\theta\|$ and some constant k. It leads to particularly simple estimates. Our preference for a general prior class does not lie here, however, for at least two reasons. Firstly, our regression model (1) transformed to this framework would necessitate a prior for θ which is data dependent in the sense that it would depend on the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Such priors have already been criticised by Lindley (1971). Secondly, the specification of a pq x pq matrix corresponding to $\mathbf{C}$ in the regression case may be somewhat daunting. It seems important to impose a fair degree of structure on any prior distributions assumed. However one feature of the Berger prior important to us is that in common with t-like priors it will produce estimates $\hat{\theta}$ satisfying.

$$\hat{\theta} = (\mathrm{I} - \mathbf{G}/\mathbf{z}^T H\mathbf{z})\ \mathbf{z} + \mathrm{o}(\|\mathbf{z}\|^{-1}) \quad \text{as } \|\mathbf{z}\| \to \infty \tag{6}$$

Let us now concentrate on the normal distribution part of the prior specification. It has been indicated in the above paragraph that highly flexible priors with general covariance structures although avoiding specific criticisms such as Rothenberg's (1963), see also Press (1972) §8.62, allow, for me at least, too much room for manoeuvre. In other contexts other authors have also advocated simplifying prior structure. For example when seeking inference concerning a dispersion matrix, Dickey, Lindley and Press (1978) suggest simple intraclass forms (see in addition Lindley, 1978).

One promising approach is to construct the prior distribution in *stages* as in the spirit of Lindley and Smith (1972). In fact, A.F.M. Smith (1971), unpublished Ph. D. Thesis, University of London, with $\Sigma = \mathbf{I}$ details priors with exchangeability between and within regressions. Suppose

$$\beta^i \sim N_p\ (\mu,\Sigma) \qquad i = 1,\ldots,q$$

where $\mu^T = (\mu_1,\ldots,\mu_p)$. This constitutes the 'between-equation' exchangeability. The 'within-equation' exchangeability is given by

$$\mu_i \sim N(\gamma,\sigma_\gamma^{\ 2}) \qquad i = 1,\ldots p$$

Further stages are envisaged but it might be natural to assume the prior distribution for γ vague, leaving just $\sigma_\gamma^{\ 2}$ and Σ (pxp) to be specified.

The approach of Brown and Zidek (1980) is somewhat different but also within the framework of Lindley and Smith. Their general covariance structure is imposed across rather than within equations. In particular if it may be assume that

$$\beta_i \sim N_q\ (\mathbf{0},\ \Gamma_\beta), \quad i = 1,\ldots p \tag{7}$$

where $\beta_1,\ldots,\beta_p$ denote the rows of β, then a class of Bayes estimates result, namely $\hat{\beta}\ (K)$ given as

$$\hat{\beta}(K) = (\mathbf{X}^\mathbf{T}\mathbf{X}\ \oplus\ \mathbf{I}_q + \mathbf{I}_p\ \oplus\ K^T)^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y} \tag{8}$$

where $K = \Gamma_\beta^{-1}\Gamma$ is a (qxq) matrix. Here $\oplus$ denotes the usual Kronecker product thought of as operating on the matrix $\mathbf{X}^T\mathbf{Y}$ row by row. More general prior assumptions allowing in (7) a non-zero mean and possible dependence of both prior mean and covariance matrix on i, $i = 1,\ldots,p$ are given in Brown & Zidek (1980). Of course the appropriateness of the priors depends on the application but (7) leading to (8) warrants special attention and perhaps deserves the name *multivariate ridge regression* since $q = 1$ corresponds to univariate ridge regression so that K (qxq) will be termed the ridge constant matrix. The estimate (8) entails the (qxq) unknown matrix K which requires estimation from the data using implicit or explicit prior information. Further justification for the use of the term 'ridge regression' will be given in Section 3.

Note how (7) hypothesises that corresponding coefficients across equations are independent and identically distributed. The simplicity of (8) makes it a favoured candidate at least after appropriate scaling of the p independent variables. This effectively allows a scalar multiplier c_i to Γ_β $i = 1,\ldots p$. One perhaps routinely important divergence from (7) is the idea that the constant vector should have a rather different prior structure. In particular a diffuse prior is traditional here (Brown, 1977).

Finally note that (8) does indeed utilise information across all equations in estimating the coefficients of any one equation.

In the next section properties of ridge estimates are examined further and some methods of estimation of K given. Particularly favoured estimates, utilised in the case study of section 5 are given by (21) and (22).

3. MULTIVARIATE RIDGE REGRESSION ESTIMATES

3.1. *Canonical form of regression model*

Following the development of Brown and Zidek (1978), let us reduce model (1) to canonical form so that properties may be readily perceived. Accordingly let

$$\mathbf{X} = \mathbf{Q}\Lambda^{1/2}\mathrm{P},\ \ \Lambda = \mathrm{diag}\ (\lambda_1,\ldots\lambda_p),\ \ \lambda_1 \geq\ldots\geq\lambda_p\geq\ \ 0$$

where the (pxp) orthogonal matrix P is such that $\mathbf{P}\mathbf{X}^T\mathbf{X}\mathbf{P}^T = \Lambda$ and the (nxp) matrix $\mathbf{Q}$ equals $\mathbf{X}\mathbf{P}^T\Lambda^{-1/2}$ so that $\mathbf{Q}^T\mathbf{Q} = \mathrm{I}_p$. Now model (1) may be expressed as

$$\mathbf{Z} = \Lambda^{1/2}\ \alpha\ +\ \epsilon^* \tag{9}$$

with $\mathbf{Z} = \mathbf{Q}^T\mathbf{Y}$, $\alpha = \mathbf{P}\beta$ and $\epsilon^* = \mathbf{Q}^T\epsilon$. Here **Q** just provides a linear reduction from n to p observations within each of the q responses. With Γ known this reduction retains the sufficient statistics for the (pxq) unknown matrix β. It results in the loss of a Wishart variable with $(n\text{-}p)$ degrees of freedom. When the case of Γ unknown is considered this Wishart variable will be utilised.

Writing $\epsilon^* = (\epsilon^{*1},\ldots,\epsilon^{*q})$ then (2) transforms to

$$E\,(\epsilon^{*j}) = \mathbf{0}, \quad \text{cov}\,(\epsilon^{*j}, \epsilon^{*\ell}) = \gamma_{j\ell}\mathbf{I}_p; \quad j,\ell = 1,\ldots q$$

so that the covariance structure is essentially unchanged. Furthermore, the prior distribution (7) is left unchanged by the transformation $\alpha = \mathbf{P}\beta$ so that

$$\alpha_i \sim N_q\,(\mathbf{0},\Gamma_\beta) \qquad i = 1,\ldots p \tag{10}$$

3.2. *The Ridge Class and its order properties*

The ridge class of estimates (8) now becomes after some manipulations

$$\hat{\alpha}_i\,(\mathbf{K}) = \hat{\alpha}_i\,\left[\mathbf{I}_q - \mathbf{B}_i\,(K)\right], \quad i = 1,\ldots,p \tag{11}$$

where $\hat{\alpha}_i$ $(1xq)$ is the least squares estimate of α_i $(1xq)$ and

$$\mathbf{B}_i\,(K) = K(\lambda_i\mathbf{I}_q + K)^{-1}$$

These shrinkage matrices satisfy the inequalities

$$\mathbf{B}_1 \leq \qquad \ldots \qquad \leq \mathbf{B}_p \tag{12}$$

where $\mathbf{A} \leq \mathbf{B}$ means **B-A** is a non-negative definite matrix. When $q = 1$ this was regarded by Thisted (1976) as an essential property of ridge shrinkage: poorly estimated coefficients (small λ_i) are shrunk most. Matrix shrinkage provided by (12) is not quite so intuitive but it does mean for example that if $\hat{\alpha}_i$

$\propto \hat{\alpha}_i$ and $\lambda_i > \lambda_j$ then

$$\hat{\alpha}_i\,(K)\,\hat{\alpha}_i\,(K)^T/\hat{\alpha}_i\hat{\alpha}_i^T \geq$$

$$\hat{\alpha}_j\,(K)\,\hat{\alpha}_j\,(K)^T/\hat{\alpha}_j\,\hat{\alpha}_j^T.$$

Thus shrinkage in relative length in particular directions of q-dimensional space strictly increases as the eigenvalues decrease provided the least squares estimates lie in the same direction.

3.3. *Choice of the Ridge Constant Matrix*

The use of a prior distribution for K will result in a posterior mean for α_i given by $E\alpha_i{}^*\,(K)$ where

$$E\hat{\alpha}_i\,(K) = \hat{\alpha}_i\,E\,\left[\mathbf{I}_q - \mathbf{B}_i\,(K)\right] \quad i = 1,\ldots p \tag{13}$$

and the expectation is with respect to the posterior distribution of K given the data. This estimate is however outside the ridge class (11). Note that when $q = 1$, $B_i\,(K) = K/\,(\lambda_i + K)$ is a concave function of K so that by Jensen's inequality

$$B_i\;(EK) \geq EB_i\;(K) \tag{14}$$

so that the member of the ridge class with K estimated by $E(K)$ shrinks more than the Bayes posterior mean estimate. The integrated ridge estimated (13) does not seem to be particularly easy to calculate and has not been considered further by us.

The literature on univariate ridge regression suggest various estimates of K, the ridge constant which may be extended to the multivariate (qxq) ridge matrix K. All fall short of the full Bayes approach mentioned in the previous paragraph. They separate into three main categories:

(i) pseudo maximum likelihood
(ii) type II maximum likelihood
(iii) empirical Bayes.

The first of these is the simplest and in that sense most applicable. Since $K = \Gamma_\beta^{-1}\,\Gamma$, Γ is known, Γ_β is the variance-covariance matrix of each α_i and these α_i have maximum likelihood estimates $\hat{\alpha}_i$, a natural estimate of K is given by

$$(\Sigma\, \hat{\alpha}\,{}_i^T\hat{\alpha}_i)^{-1}p\ \Gamma \qquad (15)$$

This extension of the univariate rule of Hoerl, Kennard and Baldwin (1975) would replace Γ by $\hat{\Gamma}$, the maximum likelihood estimate of Γ if it were unknown. Since the $\hat{\alpha}_i$ have different precisions, perhaps a further natural adaptation of the rules is

$$\hat{K} = (\Sigma\, \lambda_i\, \hat{\alpha}\,{}_i^T\hat{\alpha}_i)^{-1}(\Sigma\, \lambda_i)\Gamma \qquad (17)$$

This relative weighting of $\hat{\alpha}_i$ corresponds to a univariate estimator given by Sclove (1973), and utilised by Lawless and Wang (1976). Intuition suggests that the Hoerl-Kennard-Baldwin rule might undershrink whereas (17) might tend to overshrink.

Both (ii) and (iii) above derive from the marginal distributions of the observations given K. With (ii), this marginal distribution as a function of K, given the data, is the type II likelihood (Good, 1963) and is maximized to provide an estimate of K. The approach (iii) in the spirit of Efron and Morris (1972), looks for a function of the data which is unbiased for K with respect to this marginal distribution.

The marginal distribution of $\mathbf{Z}_i$ from model (9), (10) is q-variate normal with mean zero and covariance matrix $\Gamma\mathbf{B}_i^{-1}$ where

$$\mathbf{B}_i = (K + \lambda_i\mathbf{I}_q)^{-1}K.$$

Assuming Γ known and without loss of generality to be I_q, the type II likelihood of K is $L(K)$ where

$$L(K) = \Pi_{i=1}^{p}\ \left\{|\mathbf{B}_i|^{1/2}\exp\left(-(1/2)\,\mathbf{z}_i\,\mathbf{B}_i\,\mathbf{z}_i^T\right)\right\}$$

$$= |\Pi\,\mathbf{B}_i|^{1/2}\exp\left(-(1/2)\ \Sigma\,\mathbf{z}_i\,\mathbf{B}_i\,\mathbf{z}_i^T\right)$$

and numerical maximization of $L(K)$ gives $\hat{K}$ of (ii).

Let $w_q(\Sigma, n)$ denote a Wishart distribution with scale matrix Σ, degrees of freedom n and dimension q. From the marginal normality of $\mathbf{Z}_i$ above we know that

$$L(\mathbf{B}_i^{1/2}\ \mathbf{Z}_i^T\,\mathbf{Z}_i\,\mathbf{B}_i^{1/2}|K) = w_q(\Gamma,1)$$

Thus, in particular,

$$E\,[\,\Sigma\mathbf{B}_i^{1/2}\ \mathbf{Z}_i^T\mathbf{Z}_i\,\mathbf{B}_i^{1/2}\ |K] = p\,\Gamma$$

and if we are able to choose $K = \hat{K}$ such that $f(K)$ given by

$$f(K) = \Gamma^{-1}\ \Sigma_{i=1}^{p}\ \mathbf{B}_i^{1/2}\ \mathbf{Z}_i^T\,\mathbf{Z}_i\ \mathbf{B}_i^{1/2}\,/p$$

is the qxq identity matrix then this is the $\hat{K}$ of (iii). In the univariate case ($q = 1$) the estimate has been used by Dempster (1973) and Dempster et al (1977). In this univariate case with ridge constant K, $f(K)$ is monotonically increasing as a function of K with $f(0) = 0$ and

$f(\infty) = \sigma^{-2}\ \Sigma_{i=1}^{p}\ \lambda_i\,\hat{\alpha}_i^2/p$. Hence $f(\hat{K}) = 1$ is not achievable when $\Sigma\,\lambda_i\hat{\alpha}_i^2/p$ is less than 1 when $\hat{k} = \infty$ provides the closest solution. Evidently in the multivariate case the equation $f(K) = \mathbf{I}_q$ will have a solution if

$\Gamma^{-1}\,\Sigma\,\lambda_i\,\hat{\alpha}\,{}_i^T\hat{\alpha}_i/p - \mathbf{I}_q$ is non-negative definite and otherwise $\hat{K}^{-1} = \mathbf{0}$ provides the closest solution to $f(K) = \mathbf{I}_q$.

4. SAMPLING PROPERTIES AND DOMINANCE OF LEAST SQUARES

Brown and Zidek (1980) for Γ known, extending the univariate work of Thisted (1976), have determined sufficient conditions for estimator (11) with

$$\hat{B}_i = \mathrm{T}_i\,v_i\,w_i^i\,(c_i\,\mathbf{I}_q + \Gamma^{-1}\,\Sigma w_j^i\,\hat{\alpha}_i^T\hat{\alpha}_j)^{-1} \qquad (18)$$

with $\mathrm{T}_i > 0$, $c_i \geq 0$, $w_j^i \geq 0$ arbitrary scalars and $v_i = \lambda_i^{-1}$, to dominate least squares. Note that this is a wider class than ridge. They take as their loss function,

$$\Sigma_{i=1}^{p}\quad L_i\,(\hat{\alpha}_i^*-\alpha_i)\,\Gamma^{-1}\quad(\hat{\alpha}_i^*-\alpha_i)^T \qquad (19)$$

In this they regard $L_i = 1$ as of particular importance. Softer results may be obtained if $L_i = \lambda_i$ which arises when, as in Dempster et al (1977), the sum of squares of prediction errors at the n design points is used to measure the performance of $\hat{\beta}$. Quadratic prediction loss at m future points has been adopted by Goldstein and Brown (1978). Different m points designs lead to different L_i. Prediction at the design points $L_i = \lambda_i$ favours estimators which only slightly shrink the poorly estimated coefficients (small λ_i).

In Brown and Zidek (1979) the assumption of Γ known has been relaxed by utilising a trace ordering argument in addition to Stein's method of unbiased risk estimation (Stein, 1973). In this case the Wishart variable discarded in the canonical reduction to (19) is used replacing Γ^{-1} in (18) by $\mathbf{R}$ where $\mathbf{R}^{-1}d$ is w_q (Γ, n-p) with d an appropriately chosen scale factor. The choice $d = n$ corresponds to maximum likelihood estimation of Γ but this was not generally the preferred choice in Brown and Zidek (1979). Members of the class (18) within the multivariate ridge class have $c_i \propto v_i$ and $\mathrm{T}_i\ v_i\ w_i^i\ c_i^{-i} = 1$, with

$$\hat{K}^{-1} = \Gamma^{-1} \Sigma\ w_j\ \hat{\alpha}_j^T \hat{\alpha}_j \tag{20}$$

($\Gamma^{-1} \rightarrow \mathbf{R}$ in the unknown covariance situation). Two types of ridge estimator are given by $\hat{K}$ as in (15) (multivariate Hoerl, Kennard and Baldwin) or (16) (Multivariate Sclove). In terms of β these are given by (18) with respectively $\hat{K}$ given by

$$c\ (\hat{\beta}^T\ \hat{\beta})^{-1}\mathbf{R}^{-1} \tag{21}$$

and

$$tr\mathbf{X}^T\ \mathbf{X})\ (c/p)\ (\hat{\beta}^T\ \mathbf{X}^T\ \mathbf{X}\hat{\beta})^{-1}\mathbf{R}^{-1} \tag{22}$$

Here $c = p$ corresponds to pseudo-maximum likelihood as given by (15) and (16) whereas $c = p$-q-1 is the preferred choice in the minimax context of Brown and Zidek (1980). The latter will be distinguished from the former in subsequent description by the qualifiers 'minimax' or 'modified'. The preferred estimator of Γ via $\mathbf{R}^{-1}$ corresponds to $d = n+p+q+1$ within the minimax framework.

The multivariate ridge estimators given by (8) and $\hat{K}$ as in (21) or (22) satisfy the desired form (16) discussed earlier in connection with strength of overall prior assumptions. Thus one desired requirement has been met without explicitly expressing the complete prior distribution.

Sufficient conditions for dominance of least squares detailed in Brown and Zidek (1980), specialised to multivariate ridge estimators, determine that dominance can be achieved provided that the eigenvalue spectrum is not too wide. For example when the quadratic loss is given by $L_i = 1 \quad i = 1,\ldots,p$ the modified Hoerl-Kennard-Baldwin multivariate ridge estimator dominates least squares provided

$$(n\text{-}p)\ (p\text{-}q\text{-}1)\ v_p^{\ 2} - 2\ (n\text{-}p\text{-}2q\text{-}2)\ \{p\ \overline{v}^2 - (q\ +\ 1)\ v_p^{\ 2}\} < \mathrm{O} \tag{23}$$

with $v_i = \lambda_i^{-1}$, in the unknown Γ case. Note that this seems to run counter to the usual wisdom that ridge regression is valuable precisely when near multicollinearity is present. The fault here perhaps lies with the yardstick-least squares. Although one cannot be sure to do better than it, bolder shrinkage will pay off in most circumstances, at least as long as the prior assumptions are merely approximately valid. Further comments directed at this point are given after the election forecasting example of section 5.

5. APPLICATION TO SCOTTISH ELECTION DATA

5.1. *Description of the Data*

Although the data is only meant to illustrate the substantial improvements in performance of multivariate ridge regression over maximum likelihood, the full set of raw data employed is included so that the reader may undertake further analysis. This raw data is presented in Table 1 and consists of all 71 Scottish constituencies as defined by the two British General elections of February and October 1974. Each of the 71 constituencies is identified by its abbreviated name (e.g. "EDBR E" is Edinburgh East) and a number (from 1 to 635) denoting its order of declaration in the totality of 635 British constituencies which fielded in February 1974. The constituencies have also been ordered on this declaration order label. Thus Kilmarnock, the first in the list was the first Scottish constituency to declare but had 132 predeclarers in Britain.

Variables headed W_1, W_2, W_3, W_4, denote votes to the Conservative, Labour, Liberal, Nationalist parties in October 1974 as do those headed C, S, L, N which apply to February 1974 (in the same order). Votes for other parties have not been recorded here. Variable E is an electorate figure (February and October figures differed insignificantly) and R is a categorical variable defining region where

1 = Glasgow; 2 = Rest of Clydeside conurbation;
3 = Edinburgh; 4 = Rest of industrial centres;
5 = Highlands; 6 = Rest of Scotland.

TABLE 1

SCOTTISH GENERAL ELECTION DATA FOR FEBRUARY AND OCTOBER 1974

	R	C	S	L	N	W₁	W₂	W₃	W₄	E	Order
KILMNOCK	4	13817	23544	4878	7644	9203	22184	2508	14655	60380	133
GL CFN	1	3435	9400	982	2211	1880	9231	605	2790	25516	137
FIFE E	6	21172	6634	7766	8593	16116	7040	5247	13202	56453	170
DUMFRIES	6	21707	12739	5642	9186	18386	12558	3961	12542	61857	189
GL PROVN	1	6324	23154	0	7367	3448	20602	0	10628	54975	206
GL SHETL	1	6472	14208	0	5834	3543	13391	690	7042	38324	241
GL SPRNG	1	7452	18067	0	7672	4245	17444	865	9049	48066	264
GL CATH	1	18247	16152	0	5410	16301	14544	1058	6292	49826	276
EDBR N	3	16417	9404	5487	4550	12856	8465	3677	7681	47215	278
GL BARSC	1	9771	21035	0	8789	5004	19737	1915	12100	54700	293
AYR	4	21626	16528	0	4706	17487	14269	2611	6902	51976	314
EDBR PNT	3	18162	13560	6870	5491	14083	12826	4411	10189	54955	330
GL KELVN	1	10717	13115	0	5666	7448	11567	1735	6274	42654	341
GL MARYH	1	6625	20303	0	8920	3160	19589	1063	10171	51545	342
ANGUS S	6	20522	5721	0	15179	15249	4103	2529	17073	52275	347
EDBR E	3	14614	20163	3998	7128	10111	19669	2578	11213	57460	351
EDBR CEN	3	10393	11354	4180	4074	7176	11129	2463	6866	40956	354
EDBR S	3	18784	12403	8073	5770	14962	11736	5921	9034	56154	355
DUNDFE F	6	13371	17100	0	20066	7784	15137	1302	22120	63152	369
COAT+AIR	2	13162	24945	0	7961	7683	23034	1446	12466	59903	373
MTHWL+WS	2	11997	18310	0	7852	7069	17319	1126	12357	51506	375
BOTHWELL	2	12725	22326	5362	6710	8125	22086	4057	11138	59358	380
EDBR LTH	3	11883	12604	0	6569	8263	11708	1836	7688	39407	385
GL QPARK	1	7517	15883	0	4394	4421	14574	966	5660	38776	394
GRNOK+PG	4	7892	20565	8789	4881	4969	21279	8580	9324	62126	396
EDBR W	3	18908	10431	9189	4241	15354	10152	6606	8135	52569	397
PAISLEY	2	14723	23820	0	10455	7440	21368	3116	15778	66059	401
DUNDEE W	6	15745	22193	0	12959	8769	19480	2195	16678	63916	408
ABERDN N	6	8115	23193	6001	11337	5125	23130	3700	13509	65230	409
RUTHGLEN	2	14852	19005	0	6089	9248	17088	2424	9732	48824	411
GL POLOK	1	17684	21090	0	6584	11604	18695	2274	10441	59451	413
E KILBRD	4	15456	23424	0	13819	8513	21810	2644	19106	65799	414
GL HILHD	1	14378	7997	6644	3702	11203	8507	3596	6897	41726	415
BANFF	6	8252	1528	3121	11037	8787	1700	2059	10638	31992	417
LANRKS N	2	14664	21448	0	8187	9665	19902	1899	11561	54147	419
AYRS CEN	4	17362	23639	0	7255	11633	21188	2640	11533	59273	436
GL GOVAN	1	3049	10326	763	9783	1623	11392	444	9440	32094	438
HAMILTON	2	7977	19070	0	12692	3682	18487	1559	15155	50346	440
STIRL FG	4	12228	21685	0	17836	7186	22090	1477	20324	64362	444
DUNFMLIN	4	14791	19201	6153	8695	10611	18470	3800	13179	60680	445
GL CRAIG	1	10817	18055	0	6303	6734	16952	1728	8171	44333	449
DNBTNS C	2	9775	16439	2583	5906	6792	15837	1895	11452	49358	450
PERTH+EP	6	21167	6784	4644	12192	16544	5805	2851	17337	57646	455
RENFRW E	2	25713	10227	9588	5268	19847	9997	7015	11137	61811	457
MIDLOTHN	4	20478	32220	0	19450	11046	28652	4793	24568	89191	462
RENFRW W	4	19510	22178	5022	8394	14399	20674	3271	15374	67078	464
KIRKALDY	4	13087	22469	0	12311	7539	20688	2788	14587	60824	465
DNBTNS W	4	13638	16247	0	11144	9421	15511	2029	13697	51944	467
GALLOWAY	6	13316	3091	4643	9308	12212	2742	3181	12242	39407	469
W LOTHIN	4	11804	28112	0	21690	6086	27687	2083	24997	77527	474
BERWK+EL	6	21234	20694	0	6956	17942	20682	2811	6323	57503	475
LANARK	4	14723	16823	0	8803	9222	14948	1374	14250	48409	487
FIFE C	4	9098	24418	0	10324	5308	22400	0	14414	58403	488
ABERDN S	6	21938	18380	7447	7599	18475	18110	5018	10481	68241	490
STIRL W	4	12789	17730	0	12886	7875	16698	1865	16331	52989	492
DNBTNS E	2	19092	15416	5936	11635	15529	15122	3636	15551	61779	493
ANGUS NM	6	14288	3745	4412	6837	11835	3354	2700	9284	37604	533
ABERDN E	6	12634	2416	2727	18333	11933	3173	2232	16304	47736	552
CAITH+SD	5	5104	8574	6222	3814	4240	7941	4949	5381	28837	554
MORAY+NA	6	14279	2299	0	16046	12300	2985	2814	12667	41174	562
AYRS N+B	4	17166	10436	3832	6104	13599	10093	2224	9055	49071	573
ABERDN W	6	17256	4661	15616	6827	15111	5185	12643	9409	35341	580
AYRS S	4	10643	23093	0	6612	7402	22329	2130	7851	51330	581
STIRL EC	4	9994	18679	0	22289	5369	18657	1268	25998	62693	607
KINRS+WP	6	14356	2694	3807	6274	11034	2028	2427	10981	35237	621
ROX SL+P	6	16690	3089	25707	3953	12531	4076	20006	9178	57825	624
ARGYLL	5	12358	4027	0	15646	11036	4103	0	14967	41814	628
INVERNES	5	11680	7258	16703	7816	8922	6332	13128	11994	57527	631
ORKNY+SH	5	4186	2865	11491	0	2495	2175	9877	3025	26289	632
ROSS+CRM	5	7908	4336	4621	5037	7954	3440	1747	7291	29411	633
W ISLES	5	1042	2879	0	10079	1180	3526	789	8758	22477	635

From this raw data were generated the four response variables Y_1, Y_2, Y_3, Y_4 and seven 'independent' variables $X_1,\ldots,X_7$ where

$$X_1 = \mathrm{C/E};\ X_2 = \mathrm{S/E};\ X_3 = \mathrm{L/E};\ X_4 = \mathrm{N/E};$$

$$Y_i = W_i/\mathrm{E} - X_i,\ i = 1,\ldots,4.$$

$$X_5 = \begin{cases} 0.5 & \text{Liberal intervenes, i.e. } w_3 > 0 \text{ and } L = 0; \\ 0 & \text{otherwise;} \end{cases}$$

$$X_6 = \begin{cases} 0.5 & R = 5,6, \\ 0 & \text{otherwise;} \end{cases}$$

$$X_7 = \begin{cases} 0.5 & \text{Labour or Nationalist top party in February 1974 and } |X_2 - X_4| \leq 0.2, \\ 0 & \text{otherwise.} \end{cases}$$

The value of 0.5 employed in these three dummy variables X_5, X_6, X_7 is somewhat arbitrary but was chosen so that a priori coefficients for all seven variables would be of a similar magnitude, it being relative standardisations of different variables that is important. In most ridge regression literature standardisation is achieved by centering the independent variables and scaling so that they have constant variance (either 1 or $1/n$). The $\mathbf{X}^T\mathbf{X}$ matrix for the first 25 constituencies (after centering) is given in Table 2 where it is seen that

VARIABLE	1	2	3	4	5	6	7
1	0.21	-0.15	0.03	0.01	-0.03	0.16	0.01
2		0.19	-0.06	0.03	0.14	-0.28	-0.04
3			0.08	-0.03	-0.30	0.02	-0.04
4				0.08	0.16	0.18	0.08
5					1.56	-0.20	0.23
6						0.84	0.09
7							0.84

TABLE 2. $\mathbf{X}^T\mathbf{X}$ *matrix for seven variables and 25 observations (lower triangle as upper triangle)*

such standardisation has not been adopted here. We have taken the Bayesian attitude that these variables have names which have meanings and implications for their effect. Furthermore, in relating $\mathbf{Y}$ to $\mathbf{X}$ in model (1) a constant term is envisaged so that strictly $\mathbf{X}$ consists of 8 variables. However it was thought a priori undesirable to shrink the constant term in the same way as the other variables. It was in fact left unshrunken. It may be noted from the result of Brown (1977) that the centering of variables is not necessary in this case as long as a slightly modified form of ridge shrinkage is adopted in which $K = \mathbf{0}$ for the $q = 4$ coefficients of the constant term.

5.2 *The Problem and Criteria to Judge its Solution*

Taking the first n constituencies $n = 15, 25, 45, 65$ as data results in a nx4 matrix $\mathbf{Y}$ and a nx8 design matrix $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_7)$ where $\mathbf{X}_0$ is a vector of ones. It is desired to predict the (71-n)x4 matrix $\mathbf{W} = \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4)$ over those (71-n) remaining constituencies when their corresponding independent variables given by the (71-n) x 8 matrix $\mathbf{X}^{(0)}$ are known. To do this it is natural to use the n observations to estimate β in (1) by some method, least squares and ridge being two contenders of interest to us, and then predict the (71-n) x 4 matrix $\mathbf{Y}^{(0)}$ for the remaining constituencies by

$$\hat{\mathbf{Y}}^{(0)} = \mathbf{X}^{(0)} \hat{\beta}$$

and hence predict $\mathbf{W}$ by $(\hat{\mathrm{W}}_{ij})$ where

$$\hat{\mathrm{W}}_{ij} = (\hat{\mathrm{Y}}^{(0)}_{ij} + X^{(0)}_{ij})E_i, \quad i, \ldots, (71\text{-}n); \quad j = 1, \ldots, 4.$$

Three criteria *SD, GW, PRED* for goodness of prediction were chosen where

$$SD = \{Tr\,(\mathbf{W} - \hat{\mathbf{W}})\,(\mathbf{W} - \hat{\mathbf{W}})^T/(4\text{x}(71\text{-}n))\}^{1/2}$$

$$GW = Tr\,(\mathbf{W} - \hat{\mathbf{W}})\,\Gamma^{-1}(\mathbf{W} - \hat{\mathbf{W}})/(4\text{x}(71\text{-}n)10^8) \qquad (24)$$

PRED = ≠ incorrect predictions of winning party.

The first of these 'SD' is just the square root of the mean square of prediction errors. The criterion GW, a gamma weighted squared average penalises the prediction errors according to the ease with which they were estimated as reflected in the residual variance covariance matrix as estimated by maximum likelihood. Finally PRED accumulates scores of 1 for each of the (71-n) constituencies where the party predicted to have the largest number of votes is not in fact the party which gains the largest number of votes. This measure is very crude but has the appeal of simplicity. It does not however take any account of the closeness of a particular contest. One way to accomplish this for such a measure is to calculate the probability of winning as in Brown and Payne (1975). The criterion GW is most in line with the loss function of this paper, differing from (24) in the use of $\hat{\Gamma}$ for Γ. The diagonal matrix $D(L_1, \ldots, L_p)$ has been replaced by the pxp matrix $\mathbf{X}^{T(O)}\,\mathbf{X}^{(O)}$ in the variable space formed after canonicalising $\mathbf{X}^T\mathbf{X}$. No attempt has been made to standardise GW whereas SD has been standardised and may be thought to be the estimated standard deviation of prediction.

After centering the seven independent variables the $\mathbf{X}^T\mathbf{X}$ matrix exemplified for 25 observations, presented in table 2, has $\lambda_1 = 1.71$ and $\lambda_7 = 0.01$, a very ill conditioned matrix. We have not attempted to see whether the ridge estimators used are minimax for the prediction problem with the design matrices $\mathbf{X}$ and prediction design matrices $\mathbf{X}^{(O)}$, rather we demonstrate the size of the improvement for these particular examples as the number of data points, n, changes. For further details of election might forecasting see Brown and Payne (1975).

5.3. *Results of comparison of Maximum Likelihood and Ridge*

The Sclove rule was chosen as the main ridge contender to the maximum likelihood estimator. It was preferred to the Hoerl-Kennard-Baldwin estimator because in estimating Γ_α it paid less attention to the poorly estimated α_i. The results are given in table 3.

TABLE 3. Comparison of various estimator using criteria GW, SD, PRED with n = 15, 25, 45, data points.

	LEAST SQUARES			BAYES–SCLOVE RIDGE			MINIMAX–SCLOVE RIDGE			Modified Hoerl-Kennard-Baldwin
n	GW	SD	PRED	GW	SD	PRED	GW	SD	PRED	PRED
15	975	2135	9/56*	395	1440	6/56	571	1593	7/56	11/56
25	143	1561	7/46	70	1317	5/46	71	1314	6/56	9/46
45	70	1378	4/26	47	1259	2/26	47	1228	2/26	3/26
65	52	1180	0/6	38	1045	1/6	37	1009	0/6	0/6

* *Second figure denotes number of constituencies being predicted.*

licted.

The estimator denoted Bayes-Sclove has a divisor of $p=7$ rather than the minimax choice p-q-1 $= 2$ in the construction of $\hat{\Gamma}_\alpha$. This slightly bolder shrinker performs slightly better for this set of data with respect to the all three criteria for small n, when it really matters. Although tabulations for the Modified Hoerl-Kennard-Baldwin ridge estimator are only given for the criteria PRED it is clear that this fares considerably worse than the two Sclove rules and even worse than least squares on this rather limited basis of comparison. On the other hand, on all criteria the Sclove rules do considerably better than least squares and this is despite the wide eigenvalue structure of $\mathbf{X}^T\mathbf{X}$.

The ridge estimates depend on the least squares estimates of the parameters so that with for example as few as 10 observations $\mathbf{X}^T\mathbf{X}$ happens to have less than full rank and the estimates cannot be applied without further modification. In these situations of predicting with low information it is clear that bolder shrinkage than that of minimax shrinkers is essential. In the circumstances as described in table 3 'minimax' ridge shrinkage pays off handsomely.

ACKNOWLEDGEMENT

I am grateful to A. Hui for undertaking the computing in section 5. This computing was completed whilst the author was visiting University of British Columbia during the summer of 1978 under National Research Council of Canada, Grant No. A5241. The work, as indicated by the references, owes much to fruitful collaboration with Professor James V. Zidek.

REFERENCES

BERGER, J. (1980). A robust Generalised Bayes estimator and Confidence Region for a Multivariate Normal Mean. *Ann. Statist.* **4**, 716-61.

BROWN, P.J. (1977). Centering and scaling in ridge regression. *Technometrics,* **19**, 35-6.

BROWN, P.J. and PAYNE, C. (1975). Election night forecasting, (with discussion). *J. R. Statist. Soc. A* **138**, 463-498.

BROWN, P.J. and ZIDEK, J.V. (1980). Adaptive Multivariate Ridge Regression. *Ann. Statist.* **8**, 64-74.

— (1979). Multivariate Ridge Regression with Unknown Covariance Matrix. *Tech. Report No. 79-11,* University of British Columbia, Canada.

DAWID, A.P. (1973). Posterior expectations for large observations. *Biometrika,* **60**, 664-6.

DEMPSTER, A.P. (1973). Alternatives to least squares in multiple regression. In *Multivariate Statistical Analysis,* (D. Kabe and R.P. Gupta, eds.), 25-40. Amsterdam: North-Holland.

DEMPSTER, A.P., SCHATZOFF, M. and WERMUTH, M. (1977). A simulation study of alternatives to ordinary least squares. *J. Amer. Stat. Assoc.,* **72**, 77-106.

DICKEY, J., LINDLEY, D.V. and PRESS. S.J. (1978). Estimation of the dispersion matrix of a multivariate normal distribution. *Tech. report.*

EFRON, B. and MORRIS, C (1972). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika* **59**, 335-347.

GOLDSTEIN, M. and BROWN, P.J. (1978). Prediction with shrinkage estimator. *Math. Operationsforsch. Statist. Ser. Statistics* **9**, 1, 3-7.

GOOD, I.J. (1965). *The estimation of probabilities.* Massachussets: MIT Press.

HILL, B.M. (1974). On Coherence, inadmissibility and inference about many parameters in the theory of least squares. In *Studies in Bayesian Econometrics and Statistics* (Fienberg, S.E. and Zellner, A. eds.) 555-84. Amsterdam: North-Holland.

HOERL, A.E., KENNARD, R.W., and BALDWIN, K.F. (1975). Ridge regression: some simulations *Comm. Statist.* **4**, 105-123.

JAMES, W. and STEIN, C.. (1961). Estimation with quadratic loss. *Proc. Fourth Berk Symp. Prob. Statist.,* **1**, 361-379.

LAWLESS, J.F. and WANG, P. (1976). A simulation study of Ridge and Other Regression estimators. *Comm. Statist. A* **5**, 307-23.

LINDLEY, D.V. (1971). The Estimation of Many Parameters. In *Foundations of Statistical Inference,* (V.P. Godambe and D.A. Sprott, eds.) 435-55. Toronto: Holt, Rinehart & Winston.

— (1978). The Bayesian Approach. *Scand J. Statist.,* **5**, 1-26.

LINDLEY, D.V. and SMITH, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Statist. Soc., B,* **34**, 1-41.

PRESS, S.J. (1972). *Applied Multivariate Analysis.* Toronto: Holt, Rinehart & Winston.

RAO, C.R. (1965). *Linear statistical inference and its applications.* New York: John Wiley and Sons.

ROTHENBERG, T.J. (1963). A Bayesian Analysis of Simultaneous Equation Systems. *Tech. Report.* **6315**. Netherlands School of Economics.

SCLOVE, S.L. (1971). Improved estimation of parameters in multivariate regression. *Sankhya, Ser. A.,* **33**, 61-66.

— (1973). Least squares problems with random coefficients. *Tech. Rep.* **72**, Stanford University.

STEIN, C. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. B.* **24**, 265-296.

— (1973). Estimation of the mean of a multivariate normal distribution. In *Proc. Prague Symp. Asymp. Statist.,* 345-381.

THISTED, R.A. (1976). Ridge regression, minimax estimation and empirical Bayes methods, *Tech. Rep.* **28**, Stanford University.

ZELLNER, A. (1971). *An introduction to Bayesian Inference in Econometrics.* New York: Wiley.

ZELLNER, and VANDAELE, W. (1974). Bayes-Stein Estimators for κ-means, regression and simultaneous equation models. In *Studies in Bayesian Econometrics and Statistics,* (S.E. Fienberg and Zellner, A. eds.) 627-53. Amsterdam: North Holland.

Bayesian Inference in applied statistics

A.P. DEMPSTER
Harvard University

SUMMARY

The task of assessing posterior distributions from noisy empirical data imposes difficult requirements of modelling, computing, and assessing sensitivity to model choice. Seasonal analysis of economic time series is used to illuestrate ways of approaching such dificulties.

Keywords: COMPUTING; MODEL CHOICE; SENSITIVITY ANALYSIS; TIME SERIES.

1. INTRODUCTION

The subject of this talk is Bayesian inference: what it is useful for, limitations on its applicability, and how it can shape the way an applied statistician approaches the study and analysis of empirical phenomena. The purpose is to describe how Bayesian inference ought to be implemented. The views expressed are personal opinions in the sense that they convey a set of principles which I attempt to follow as an applied statistician.

The objective of Bayesian inference is to quantify uncertain knowledge about a set of unknown quantities in terms of a posterior distribution of those unknowns. The method is to specify a joint probability distribution of a set observables and unobserved quantities, and thence obtain the required posterior distribution by conditioning on the values of the observables. The audience is assumed to be familiar with common statistical models used in Bayesian inference.

I believe that applied statistics draws on varied patterns of reasoning, both nonprobabilistic and probabilistic, and that Bayesian inference provides only a part of the latter category. For example, a procedure such as least squares may be used to fit a linear model to data, with the intention of exhibiting structure in the data, and if diagnostic plots indicate a good fit the model may be adequate to answer all the questions arising in the application, without any use of probability.

The degree of fit of the model to data can be assessed, in a sense which needs careful explanation, by means of tail area significance tests, which are probabilistic but notBayesian. Fitting a model to data does imply that the values of certain unknowns are provisionally estimated. When the application requires a more refined quantification of the uncertainty concerning these unknowns, then Bayesian inference becomes appropriate I believe that sampling distributions are irrelevant to such refined estimation and that Bayesian inference should be regarded at present as the only generally applicable, although not entirely satisfactory, mode of analysis.

I take the view that Bayesian inference is logically separable from decision analysis. It is true that the main uses of Bayesian inference are technological, such as designing a sample to meet specifications of expected accuracy or evaluating expected gains and losses of strategies available to a decisionmaker. It is also true that the effort which an applied statistician puts into the demanding task of determining a posterior distribution, or more often a set of alternative plausible determinations, should depend on the scientific or technological application, because the statistician should wish to direct effort at those refinements of posterior knowledge most useful to his client. But the client's utility function is generally very different from that of the professional statistician. In this paper, I will be illustrating the schema by which applied statisticians develop posterior distributions, and I believe it will be obvious that these schema have no logical the to any decision problem.

The plan of the paper is to discuss issues which arise in applied statistics, by making use of a current effort to develop techniques for the analysis of monthly economic time series. The specific project is being sponsored jointly by the American Statistical Association and the Bureau of the Census in Washington, D.C. It is too soon to report results, so my paper is about approaches. How well the approaches work must be reported later.

The applied problem is described in §2. The most demanding part of Bayesian inference is largely nonBayesian, namely the task of developing the overall probability model with enough detail to permit Bayesian inference. Some key issues in the modelling process are reviewed in §3. I believe that computing is much more important to Bayesian analysis than is the study of analytical properties of models. Computational aspects of the proposed time series analysis are discussed in §4. Methodological aspects of the Bayesian inference are set out in §5. Finally, implications of my position for future statistical research are explored in §6.

2. THE APPLIED PROBLEM

The problem of seasonal adjustment of economic time series has returned to center stage in recent ywars both among statisticians and economists

(Mandelbrot, 1972; Zellner, 1978) in part because the economic dislocations of 1973-1974 put severe strains on the existing official method of seasonal adjustment, namely, the X-11 program developed at the Bureau of the Census about 15 years ago by Julius Shiskin. At present the Federal Reserve Board has a review committee studying the problem, and the Bureau of the Census has two research efforts underway, one being the ASA/Census Project of which I am a member*. This is not the place to review the diverse approaches which have been proposed for seasonal analysis. Instead, I will present a simplified personal view leading to some questions for which Bayesian answers may be sought.

Hundreds of raw time series values are collected each month, such as dollar volumes of various components of retail sales, or employment statistics by age, sex, race, and region. These hundreds could easily be made into thousands by finer disaggregation. Most reported series are seasonally adjusted before publication. Is there such a thing as a good or valid method of seasonal adjustment?.

Statistical methods can only be evaluated in relation to their purposes. I see two main purposes of seasonal analysis. The first is simply to find and describe patterns of variability detectable in the data. Examples include long slow swings often informally labelled trends and cycles, abrupt changes attributable to strikes or other special circumstances, and monthly patterns which visibly repeat from year to year. A bit more difficult to detect are trading day effects which produce small predictable changes from year to year in the averages of each month due to changes in the composition of the month in terms of days of the week. Similarly, holiday effects are produced when a holiday such as Easter shifts between March and April. Finally, there are slow changes in the month effects, trading day effects, and holiday effects. These slow changes are generally difficult to describe and quantify. Applied statisticians have traditionally provided tools helpful in the identification and display of patterns such as those described in this paragraph.

The second purpose of seasonal analysis is to provide policy analysts with useful information. Usually the statistician is asked to produce an estimated series with seasonal and calendar effects removed, in order that the underlying measure of economic performance can be tracked more accurately. Indicators of rate of change, and predictions of turning points in the economy are desirable. I believe that this second purpose can only be met through posterior probability distributions of rates of changes and future values of nonseasonal components of time series. By and large the statistical community has offered only traditional descriptive statistics, and the user community has learned to live with what the statisticians give them. Both groups should raise their expectations from the statisticial profession.

[*] The other members are William P. Cleveland, Mark Abrahams, Jessica Pollner, and Joseph Stith.

A quick solution cannot be expected. Satisfying Bayesian formulations do not come easily. A particular conceptual problem which troubles Bayesian seasonal analysis concerns whether or not there is such a thing in the real world as the deseasonalized series corresponding to a given real world seasonal series. Obviously, it can make little sense to quantify uncertainty about a set of unknowns if these unknowns do not exist. The usual way to establish the credibility of an unknown is to specify how to find its value given unlimited resources. Thus, if seasonal effects remained constant over time, then observing a time series for many years would determine them arbitrarily closely, but common wisdom holds that the seasonal components are themselves somewhat random from year to year and cannot therefore be validated by replication, I believe that the problem is less severe in practice than in principle, because seasonal effects are those which receive power from spectral peaks close to the seasonal frequencies, and such peaks can be reasonably well identified empirically.

As indicated later, I believe we are some distance away from viable models for highly multivariate dynamic systems, such as the many components of an aggregate series, or the much more elaborate systems which economists profess to represent by simultaneous linear causal models. Still, I believe it is possible to do useful work related to the seasonal adjustment of single time series, and to make this work illustrate the importance of Bayesian thinking.

3. MODELLING

There is no doubt that the objective of Bayesian inference imposes difficult standards on the applied statistician's modelling task. The model must be sufficiently detailed to capture the important features of the system being modelled, and sufficiently complete in its assignment of prior probabilities that the requires posterior probabilities or expectations are mathematically defined. In addition, it must be feasible computationally to obtain numerical values for these posterior probabilities or expectations.

I see three major sources which contribute to the model-building enterprise. The most fundamental is understanding of the current state of scientific knowledge in the area involved, so that key factors are incorporated and the probability models are in accord with accepted knowledge. A second source is knowledge of mathematical structures and of how these structures are used to represent and organize thinking about empirical phenomena. The third source is data analysis, both exploratory and confirmatory, since acceptable models generally result from an iterative process of trial and error aimed at satisfactory conformity with known empirical facts. Such data analysis is the bread and butter of the statistical profession, as opposed to the rarer type of data analysis which seeks to uncover unsuspected scientific facts. The goal of Bayesian inference helps to focus data analysis on those questions which are critical to the credibility of the Bayesian inference, such as failures of normality assumptions. Otherwise much of the large and growing body of data analysis techniques strikes me as aimless.

Turning now to modelling economic time series, the requirement which I believe to be most fundamental cannot be met at present. Because the basic time series are aggregates, I believe they should be represented by linear models *in the original units*. Unfortunately, the only linear models leading to computationally viable Bayesian inferences are Gaussian linear models which do not conform to the data in original units. Hence the confused debates over additive models, multiplicative models, and combinations thereof. Econometricians usually transform the original scales to logarithms in order to obtain credible fit to Gaussian linear models. I believe the results are often acceptable for analysing single time series, but for multiple time series analysis logging is very awkward because the sum of log normal random variables is not log normal. Ultimately we will need Bayesian linear model theory based on stable distributions. Another problem with multiple time series analysis is the rapid proliferation of parameters, a problem which is already troublesome in the univariate case, and which will require hidden factor models to capture the hoped-for simplicity. The plan therefore is to develop Gausian linear models for transformed single time series, while adopting *ad hoc* procedures to allow for the inevitable stragglers in the tails. The deeper modelling required for a satisfactory treatment of more general econometric analysis is put aside for future research.

It is instructive to compare three approaches which have been suggested for time series modelling. To bring out the essential points I will start by supposing there are no seasonal or calendar effects. The most widely practiced approach is to fit autoregressive integrated moving average (ARIMA) models in the manner of Box and Jenkins (1976). Data analysis is used to assign a set of dimensions (p,d,q) whose meaning is that the dth order differences are assigned an autoregressive moving average (ARMA) model with p autoregressive parameters and q moving average parameters. A fairly typical model would be a (1,1,2) model which can be denoted by

$$(\mathrm{I}-\phi_1\mathrm{B})\,(\mathrm{I}-\mathrm{B})\mathrm{Y}_t = (\mathrm{I}-\theta_1\mathrm{B}-\theta_2\mathrm{B}^2)\,a_t \qquad (3.1)$$

where Y_t for $t=1,2,\ldots$ denotes the original time series, B denotes the backshift operator such that $BX_t=X_{t-1}$, $B^2X_t=X_{t-2}$, etc., the AR and MA parameters are denoted by ϕ_1 and (θ_1,θ_2), respectively, and a_t denotes a Gaussian white noise driver with variance σ^2.

A second approach advocated by Parzen (1977) is to fit only AR terms in the model permitting p to be fairly large, say $p=15$ or $p=30$, to secure adequate fit. Various options are to limit the coefficients $\phi_1,\phi_2,\ldots,\phi_p$ so that the Y_t process is stationary, or to introduce a special limited form of non-stationarity by using a factor of the form $(I\text{-}B)^d$, or to place no restrictions on the AR parameters so that explosive nonstationarity is permitted.

A third approach advocated by Mandelbrot (1972) is to give fractional Gaussian processes a key role in the modelling. The basic model has the form

$$(\mathrm{I}-\mathrm{B})\mathrm{Y}_t = a_t \qquad (3.2)$$

where a_t is no longer Gaussian white noise but a generalization called fractional Gaussian noise which is a stationary Gaussian process with autocovariance function

$$c(t) = \tfrac{1}{2}\sigma^2[(t+1)^{2H}-2t^{2H}+(t-1)^{2H}] \qquad (3.3)$$

where σ^2 is the variance and H is a fractional parameter on $O<H<1$. When $H=\tfrac{1}{2}$ the a_t process is white noise. The family (3.2) can be extended in various ways, for example, by multiplying the spectral density $f(\lambda)$ corresponding to (3.3) by a smooth spectral density $h(\lambda)$ to change a_t to a stationary process with spectral density $f(\lambda)h(\lambda)$, or by changing the order of differencing $d=1$ in (3.2) to $d=0$ or $d=2$.

What are some arguments for and against these different modelling strategies? The basic argument for ARIMA is parsimony. For example, a (1,1,2) model is completely described by the choice $d=1$ and the four real-valued parameters ϕ_1,θ_1,θ_2, and σ^2. A successful fit of such a model to and observed time series Y_t obtained when the residuals $\hat{a}_t$ implied by the fitted model cannot be told from white noise in senses made precise by Box and Jenkins (1976). Such parsimonious fit can often be obtained. I argue, however, that while parsimonious fit is a worthy goal if the purpose of statistical analysis is to exhibit detectable patterns in the data, such fit is potentially misleading if the ultimate goal is realistic Bayesian inference. The reason simply is that probability models underlying Bayesian inference should reflect plausible prior knowledge, and there can rarely be prior knowledge which fixes $(p,d,q) = (1,1,2)$. The situation can be redeemed by showing that more realistic prior assessments do not have meaningful effects on the final Bayesian inferences, but such demonstrations in effect require practical performance of the more refined Bayesian analyses. My criticism is not that certain alternative analyses are not routinely performed, as when extra terms in the model are shown to affect little. The key distinction is between *ad hoc* parsimony which mechanically stops at $(p,d,q,) = (1,1,2)$ because the data are inadequate to detect further signals in noise, and true parsimony which postulates simple structure in nature and seeks corresponding prior assessments that may prove effective in the long run. My concern is that continuing lack of attention to this distinction may have a cumulative destructive effect on the credibility of Bayesian analysis.

The second approach rejects parsimony and conforms to the dictum which I first recall hearing from Jimmie Savage to make the model "as big as an elephant". The reason is to scan the data for a wide range of possible messages. A key trouble with the elephant principle is that it conflicts with another principle which Jimmie and most Bayesian statisticians have considered a basic reason for the feasibility of Bayesian inference, namely, the principle of precise measurement, which holds that the nonempirical component of the prior probability assessment is less critical than the empirically checkable component in data rich situations. I believe that when there are many parameters, the likelihood is rarely peaked enough to overwhelm the prior, so that simple alternative prior distributions which assign plausible finite variances to the parameters yield substantially different inferences from flat priors. I find the use of criterion decision rules to produce an automatic AR dimension p, followed by an automatic flat prior analysis of the p parameters, to be highly contradictory to the spirit of Bayesian analysis. I am more sympathetic to analyses which generate genuine priors for many parameters by means of hierarchical models, as advocated by Good (1965) or Lindley and Smith (1972), but I question whether the full story is in on the strengths and weaknesses of such models.

The reason why I am attracted to Mandelbrot's proposal is that it hypothecates a single parameter H to capture a real world phenomenon which can only be approximated by a substantial number of physically meaningless AR parameters. If successful, the hypothesis will exhibit genuine scientific parsimony, as opposed to the dubious statistical parsimony which accepts simple *ad hoc* models primarily because they cannot be empirically disproved with available data. Simulations of the model (3.2) produce hypothetical time series whose long swings are reminiscent of the trends and cycles of real economic time series, especially for H in the range .7 to .9. The generator a_t governed by (3.3) has spectral density proportional to λ^{1-2H} near $\lambda = 0$, which suggests that the nonstationary Y_t in (3.2) has a nonintegrable spectral "density" proportional to λ^{-1-2H} near $\lambda = 0$, and it is this smooth progression across a class of nonstationary models as H varies which gives the models the capability of representing and interesting range of low frequency behavior.

Such spikes can be simulated close to $\lambda = 0$, but not actually at $\lambda = 0$, by using increasing numbers of AR parameters, but should one choose to spend parameters this way if a single plausible parameter will do?

Having decided to build models with fractional power spectral peaks, the main lines of my modelling strategy are set. Seasonal components in the model will also be hypothesized to have fractional power spikes in the spectrum at the seasonal frequencies, in order to meet the scientifically real need for a simple model capable of representing slow drifts in seasonal patterns. Further details are sketched in §4.

The reason for the elaborations of §3 has been my wish to exhibit the kind of science-oriented but nonBayesian deliberations which necessarily precede serious Bayesian inference from statistical data.

4. COMPUTING

The application of inference techniques is held back by conceptual factors and computational factors. I believe that Bayesian inference is conceptually much more straightforward than nonBayesian inference, one reason being that Bayesian inference has a unified methodology for coping with nuisance parameters, whereas nonBayesian inference has only a multiplicity of *ad hoc* rules. Hence, I believe that the major barrier to much more widespread application of Bayesian methods is computational. Being less limited conceptually, Bayesian statisticians should tackle the computing problems associated with complex data sets and correspondingly complex models. The development of the field depends heavily on the preparation of effective computer programs.

These programs include methods for exploring data, exploring models, and fitting models to data, but the central computing problem of Bayesian inference is the problem of computing posterior probabilities and expectations from high dimensional distributions. In the computer era, it is not sufficient, and often spurious, for statistical theorists to express their techniques in terms of formulas and equations, since the basic requirement is to provide feasible algorithms.

With Gaussian linear models, the situation is much helped, because posterior distributions are multivariate normal and the marginal posterior distribution of single components is normal. Hence many posterior calculations depend on numerical linear algebra rather than numerical multiple integration. With most other models, mathematical analysis is largely intractable, although expansions around Gaussian models can be helpful. In the end, I expect that numerical sampling techniques will provide most of our useful posterior calculations.

For concreteness, I now spell out some details of Gaussian linear model posterior analysis and its applications to time series modelling. The basic form of the model is

$$\mathbf{Y} = \mathbf{X}\beta \qquad (4.1)$$

where β is a $px1$ vector of unknowns which govern the system, $\mathbf{X}$ is a known nxp linear transformation matrix, and $\mathbf{Y}$ is an $nx1$ vector of observations. If β has a multivariate normal prior distribution, and $\mathbf{X}$ has rank n, then the model (4.1) asserts that the observed $\mathbf{Y}$ conditions β by forcing it to lie in an $(n\text{-}p)$-dimensional subspace of its p-dimensional space, and hence the posterior distribution of β is simply the restriction of the prior normal to the subspace, which is again multivariate normal.

The model (4.1) is more general than models which look more general. For example, the familiar multiple regression model $\mathbf{Y} = \mathbf{Y}\beta + \mathbf{e}$ is just a special case of (4.1) when $\mathbf{e}$ is appended to β and a corresponding $\mathbf{I}$ is appended to $\mathbf{X}$. The mean of the normal prior of β may be included as another term in the model and hence the prior mean of the general model can be taken to be a vector of zeros. The prior distribution of β can be stretched infinitely far along r dimensions, thus incorporating flat priors for compounds of β thought to include components with diffuse prior knowledge, while the posterior remains finite normal as long as the r-space intersects the n-space determined by $\mathbf{Y}$ in a space of zero dimension.

In most applications, it cannot be assumed that the prior covariance matrix Σ of β is known, so that a second level of prior distribution is required for Σ. It turns out, however, that the computations for Σ known are central to working with many models whose Σ is only partially known. In the remainder of §4 Σ is taken to be known, while §5 the case of unknown Σ is taken up.

To exhibit the flexibility of the model (4.1), I now describe a time series modelling effort leading to such models. The first model will have four subcomponents. That is, the elements of β will be partitioned, and the rows of $\mathbf{X}$ correspondingly partitioned, so the model appears in the form

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3 + \mathbf{X}_4\beta_4 \qquad (4.2)$$

I take the components of $\mathbf{Y}$ to be a time series Y_t for $t = 1,2,\ldots n$ and represent

$$Y_t = n_t + s_t \qquad (4.3)$$

where n_t is a time series of nonseasonal components and s_t is a time series of seasonal components.

The model (3.2) is adopted for n_t, namely,

$$(I\text{-}B)n_t = a_t \qquad (4.4)$$

where a_t is a stationary Gaussian process with spectral density of the form $f(\lambda)h(\lambda)$ described in §3. Operationally, the model (4.4) requires value n_0, and then is generated by

$$\begin{aligned} n_1 &= n_0 + a_1 \\ n_2 &= n_1 + a_2 = n_0 + a_1 + a_2 \\ n_3 &= n_2 + a_3 = n_0 + a_1 + a_2 + a_3 \\ &\text{etc.} \end{aligned} \qquad (4.5)$$

It is convenient to represent the vector of n_t as $\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$ where β_1 has the single element n_0, and β_2 is the vector $(a_1, a_2, \ldots, a_n)'$, while

$$\mathbf{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \quad \text{and } \mathbf{X}_2 = \begin{bmatrix} 100 \ldots 0 \\ 110 \ldots 0 \\ 111 \ldots 0 \\ \ldots \\ \ldots \\ \ldots \\ 111 \ldots 1 \end{bmatrix} \qquad (4.6)$$

The prior for n_0 is $N(0,\sigma_n^2)$ where σ_n^2 is a fairly large number whose exact value is unimportant, and the prior covariance for $a_1, a_2, \ldots a_n$ is that implied by the spectral density $f(\lambda)\ h(\lambda)$.

The model adopted for s_t is a seasonal analogue of (4.4), namely,

$$(\mathrm{I\text{-}B}^{12})s_t = b_t, \qquad (4.7)$$

assuming monthly data, where b_t is a stationary Gaussian process with spectral density proportional to $(\lambda\text{-}\lambda_i)^{1\text{-}2H}$ near the seasonal frequencies $\lambda_i = (i/12)$ for $i = 1,2,3,\ldots,6$. Again, the model (4.7) requires starting values, namely, $s_{-11}, s_{-10}, \ldots, s_0$ and the s_t are generated by

$$\begin{aligned} s_1 &= s_{-11} + b_1 \\ s_2 &= s_{-10} + b_2 \\ &\cdot \\ s_{13} &= s_{-11} + b_1 + b_{13} \\ s_{14} &= s_{-10} + b_2 + b_{14} \\ &\text{etc.} \end{aligned} \qquad (4.8)$$

The vector $s_1, s_2, \ldots, s_n$ is now represented as $\mathbf{X}_3\beta_3 + \mathbf{X}_4\beta_4$, where $\mathbf{X}_3\beta_3$ specifies the component depending on $\beta_3 = (s_{-11}, s_{-10}, \ldots, s_0)^t$ and $\mathbf{X}_4\beta_4$ the component depending on $\beta_4 = (b_1, b_2, \ldots, b_n)^t$. The components of β_3 are assumed independent $N(0, \sigma_s^2)$ *a priori* where σ_s^2 represents the variance of the initial seasonal pattern β_3, and β_4 represents the wandering seasonal which typically has small variance coming mainly from close to the seasonal frequencies.

More general models are easily constructed by adding terms to (4.2). For example, initial trading day effects, and wandering trading day effects provide two more terms. The error structure of the series $\mathbf{Y}$ can be assayed, and a further term can be added for sampling and/or other measurement errors. And so forth.

A computational strategy will now be sketched for (4.2). Supposing that the β_i are independent $N(\mathbf{0}, \Sigma_i)$, the joint distribution of $\mathbf{Y}$, β_1, β_2, β_3, β_4 is multivariate normal with zero mean vector and covariance matrix

$$\begin{bmatrix} \Sigma_{i=1}^{4}\mathbf{X}_i\Sigma_i\mathbf{X}_i^T & \mathbf{X}_1\Sigma_1 & \mathbf{X}_2\Sigma_2 & \mathbf{X}_3\Sigma_3 & \mathbf{X}_4\Sigma_4 \\ \Sigma_1\mathbf{X}_1^T & \Sigma_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \Sigma_2\mathbf{X}_2^T & \mathbf{0} & \Sigma_2 & \mathbf{0} & \mathbf{0} \\ \Sigma_3\mathbf{X}_3^T & \mathbf{0} & \mathbf{0} & \Sigma_3 & \mathbf{0} \\ \Sigma_4\mathbf{X}_4^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma_4 \end{bmatrix} \tag{4.9}$$

New coordinates $\mathbf{U}, \gamma_1, \gamma_2, \gamma_3, \gamma_4$ are introduced which are orthonormal, the transformations from the new to old coordinates being provided by the triangular square roots of the five diagonal covariance matrices in (4.9), as produced, for example, by the familiar Cholesky algorithm. In terms of the new coordinates, the covariance matrix (4.9) takes the form

$$\begin{bmatrix} \mathbf{I} & \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \mathbf{B}_4 \\ \mathbf{B}_1^T & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_2^T & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_3^T & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{B}_4^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{4.10}$$

whence $\mathbf{B}_i$ are trivially seen to be the filters which provide the posterior means of the γ_i given $\mathbf{U}$, while the posterior covariances of the γ_i are given by $\mathbf{I}$-$\mathbf{B}_i^T\mathbf{B}_i$. Transformation back to the original β_1, β_2, β_3, β_4 is via the triangular square roots of the Σ_i. The reason for including the orthonormalization of the β_i in the computations is that the estimated γ_i are interpretable as residuals and are used in iterative modelling.

When n_t and s_t in (4.1) are replaced by posterior means, the result is an estimated decomposition of the observed time series into seasonal and nonseasonal components. The underlying idea is widely known, was ably exposited long ago by Whittle (1963), and more recently was used by Cleveland and Tiao (1976). I hope to facilitate applications of the theory by pointing out how simple the associated computational procedures are, and by developing programs to implement these procedures.

5. BAYESIAN INFERENCE

Reported Bayesian inferences should be defensible as reflecting considered prior assessments of the real world system under study. Since available prior knowledge is rarely adequate to support a single assessment, a range of alternative analyses should be prepared, and should be reported if they affect the end uses of the inference process. The client can then make a final prior assessment with some understanding of what rides on the choice.

A model should not be accepted on the basis that it cannot be negated from available data without a search for alternatives of comparable prior plausibility, because the inability to reject does not guarantee that similarly unrejectable alternatives will not lead to importantly different end uses. The problem is most apparent regarding the traditional "prior distribution" of parametric Bayesian statistics, since such priors are often unaccompanied by any data which might limit the prior. I am arguing against passive acceptance of any prior assessment, and especially of so-called "informationless" priors.

Admittedly my prescriptions are vague and unsatisfactory, since in the end choices must be made or the possibility of refined posterior assessments must be foregone. The best hope is for statisticians to develop a set of evolving professional standards, based on wide experience, and sufficient unto the day. New insights, new computations, and new feedback from clients should be expected gradually to bring new practices.

In the case of Gaussian linear models, my guidelines call for reanalyses under different transformations and different rules for identifying and coping with outliers. Eventually, nonGaussian linear model analysis should become computationally feasible.

Within the Gaussian framework, modelling efforts generally come down in the end to a need to estimate parameters for which prior knowledge remains vague across ranges which affect end uses. For example, the stochastic component of the nonseasonal model (4.4) certainly has a unknown scale factor and a unknown fractional H, and may have a few more parameters to shape the spectral density factor $h(\lambda)$. Similar remarks apply to the stochastic seasonal model (4.7).

I wish to compare three attitudes which can be taken to parameter estimation. The simplest approach is to compute point estimates of the parameter values, and adopt the Gaussian model with the estimates substituted for the unknown values. The second approach is to adopt continuous prior densities with simple analytic forms, and to repeat the analysis with several choices to assess sensitivity. The third approach is to compute the Gaussian posteriors and the associated likelihoods at an array of points in the parameter space, and then require the user to specify a final choice of prior weights over the array, if such a choice is needed.

The first strategy if often used because it is easy to implement. Its validity depends on the scale of the posterior distribution of the parameter values being reasonably small relative to the scale of changes in parameter values required to produce important changes in end uses. The posterior mean of β calculated at $\hat{\Sigma}$ should not differ greatly from the posterior mean of β calculated by averaging over a posterior distribution of Σ, but the posterior covariance of β calculated at $\hat{\Sigma}$ will generally understate a more refined assessment of posterior variability. Hence, for example, the first method would typically lead to an overly optimistic posterior probability that a component of β resides within desirable tolerances.

The second strategy requires a quantum increase in computation. For example, if the variances are given inverse gamma priors, then the posterior density of β given H is the product of multivariate t factors, from which posterior probabilities could be computed at considerable expense, at least in terms of algorithm development, and still the posterior distribution of H would need to be found by numerical quadrature. Even assuming feasible computations, there are potential disadvantages. The client does not automatically receive direct information about the sensitivity of end uses to value changes in the parameters themselves. A separate evaluation of the sensitivity would be required, because low sensitivity would imply that the full Bayesian analysis is not needed. Moreover, if sensitivity is exhibited, then the client finds that a choice must be made in a space of hyperparameters which I find difficult to relate to personal experience.

The feasibility of the third method depends on the number of parameters and the design of the array of points where the Gaussian posterior computations are performed,. Again, a study of sensitivity is required, but here it can be part of a sequential process of array design. If sensitivity is low relative to accuracy of parameter determination by the data, then a simple grid will do. It appears that the third approach is a sensible starting place in practice.

6. IMPLICATIONS FOR STATISTICS

The future of Bayesian applied statistics clearly depends on practitioners with sound tools for complex examples, emphasizing realistic models and correspondingly tailored analyses. Statistical research should become much more oriented to eliminating the computational barriers to the widespread practice of Bayesian analysis. Academic research needs increasingly to organize itself into teams able to cope with expanding technological possibilities. Statistical science may then be freed from its excessive dependence on theoretical insights and studies, opening the way to more theories which better match real needs. I hope also that the statistical profession may win a place of leadership in applied fields such as medicine and policy analysis where current practices fall far short of achievable standards.

ACKNOWLEDGEMENT

Work supported in part by NSF grant MCS77-277119 and in part by the ASA/Census Seasonal Adjustment Project of the American Statistical Association.

REFERENCES

BOX, G. E. P and JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control* (Revised Edition). San Francisco: Holden-Day.

CLEVELAND, W.P. and TIAO, G.C. (1976). Decomposition of Seasonal Time Series: A Model for the Census X-11 Program. *J. Amer. Statist. Assoc.* **71**, 581-587.

GOOD, I.J. (1965). *The Estimation of Probabilities*. Cambridge, Mass. The M.I.T. Press.

LINDLEY, D.V. and SMITH, A.F.M. (1972). Bayes Estimates for the Linear Model. *J. Roy. Statist. Soc. Ser. B* **34**, 1-41.

MANDELBROT, B.B. (1972). Statistical Methodology for Nonperiodic Cycles: From the Covariance to R/S Analysis. *Annals of Economic and Social Measurement* **1**, 257-288.

PARZEN, (1977). Multiples Time Series: Determining the Order of Approximating Autoregressive Schemes. In *Multivariate Analysis IV,* (P.R. Krishnaiah, ed.) 283-295. Amsterdam: North Holland.

WHITTLE, P. (1963). *Prediction and Regulation*. Princeton: Van Nostrand.

ZELLNER, A. (ed.) (1978). *Seasonal Analyses of Economic Time Series*. Economic Research Report, **ER-1**, U.S. Bureau of the Census, Washington, DC.

DISCUSSION

P.J. HARRISON (*University of Warwick*):

In relation to Dr. Brown's paper, I would recall that this morning George Barnard mentioned the comparison done by I.C.I. Ltd. between Ridge Regression and ordinary Least Squares Regression. The recommendation arising from that study was to continue to use the latter.

Now one has to be very careful in making comparisons. Most of us will have suffered from what we would regard as totally unjustified comparisons. For example, consider the comparison of a sledge with a car as a means of transport. In the Artic the sledge with its husky dogs would tend to win whereas on a motorway in Spain the car clearly wins. But perhaps we are on the sea when neither a sledge nor a car is particularly useful! Consequently the major question about the current comparison concerns its relevance. Are we on ice, road or sea?

Multivariate analysis is always worrying since its successful application demands great care. In Phil Brown's example my worries are about robustness and model adequacy; about

(i) the global linearity with fixed coefficients over time;
(ii) associated Normality;
(iii) the structure of the variance matrix, particularly since the data are roughly proportions.

So by what standard can we judge the validity of the comparison? Examples give us a good opportunity for assessment and Phil is courageous enough to give us his data. Taking his specific election example, since one of the main purposes of election forecasting is to forecast as soon as possible the number of seats which will be won by each party, I looked at the 'Pred' comparison. Remembering that October 1974 and February 1974 are not far apart in time, I first postulated a 'no change' model M_0:-

'Each Party will retain a seat previously held'.

Looking at my table 1, this is seen to outperform all Phil's models as given in his table 3. Since M_o may be interpreted as a redundant ordinary Bayes regression model with an unshakeable prior, Phil's conclusions are clearly questionable if based on such examples.

TABLE 1

A comparison of various estimators using Pred.
Showing a number of incorrect forecasts for varying n.

Number of results (n)	Least sq.	Mod H-K	Min Rid	Bayes S. Rid	M_o	M_1	M_2
0					5	5	5
15	9	11	7	6	4	4	1
25	7	9	6	5	4	4	1
45	4	3	2	2	3	2	1
65	0	0	0	1	0	1	1

We may investigate the comparison further using a very simple model in which we write for seat k

$$R_{ijk} = \theta_{ij} r_{ijk} + U_{ijk}$$

where party i won the seat in February and r_{ijk} is the number of votes then cast for party j relative to those cast for party i. R_{ijk} is the corresponding quantity for October. θ_{ij} is an unknown regression coefficient and U_{ijk} a Normal random variable with zero mean and here the variance is inappropriately taken as a constant V.

Thus the model can be written over all the constituencies as

$$\mathbf{R} = \mathbf{r}\theta + \mathbf{U} \qquad\qquad \mathbf{U} \sim N(\mathbf{0};\ \text{diag}\ (\mathrm{V}))$$

where **R** and **V** are the vectors of all the meaningful R_{ijk}'s and U_{ijk}'s, **r** is a matrix with only one appropriate non zero quantity in each row and θ is the column vector of the θ_{ij}'s. For model M_1 we will take an exchangeable ignorance prior structure at time $t=0$ as

$$(\theta\,|\,\mathrm{t}=0) \sim N\,[\mathbf{1};\ \text{diag}\ 10^{100}]$$

Thus M_1 effectively performs independent least squares regressions in sequentially estimating each of the elements of θ. At any time the Pred forecast for a winner of a constituency seat is that party with the highest expected proportion of votes. The performance of M_1 is given in Table 1.

The particular purpose of this election was for the Labour Party, who held a small majority of seats, to go to the country and obtain an increased majority. Consequently rather than the ignorance prior on θ of model M_1 it could be argued that although such a prior would be suitable for seats held by parties other than the Labour Party, there was a strong priori argument to say that Labour would retain seats it previously held. If this is so then perhaps a more realistic prior for θ would have been

$$(\theta\,|\,\mathrm{t}=0) \sim N\left[\begin{pmatrix}\mathbf{1}\\ \mathbf{1}\end{pmatrix};\quad \begin{pmatrix}\text{diag}10^{100} & \mathbf{0}\\ \mathbf{0} & \text{diag}\,\epsilon\end{pmatrix}\right]$$

where the lower block only relates to seats held in February by Labour and where ϵ is small. With such a model M_2, the performance on Pred is extremely good as shown in the table and again outperforms all others with a great deal to spare.

Whether or not one accepts model M_2, the performance of the elementary models M_0 and M_1, and the results of the I.C.I. study seriously question the authors conclusions and stress the need for great care in making comparisons of technique particularly when one of them appears to lack robustness to variation in its assumptions.

There has been little time to study Professor Dempsters paper and I will just consider a few points. He asks is there such a thing as a good or valid method of seasonal adjustment? I suppose that in order to answer this we need to be very clear what we mean by seasonality and thus I would ask what does it mean?

In his case, at the Bureau of Census, if a deseasonalised series of official statistics is to be produced for many anonymous users with a variety of utilities, some of them are going to want to know what has been filtered out of the data and what has not.

For example in my own implementation of short-term forecasting situations in which we use structured state-space or Dynamic Linear Models we discuss with users what will and what will not be parameterised and what it will mean. This is particularly important with respect to seasonality. If we are considering monthly temperatures in Central England (Box-Jenkins) then we might be happy to describe the seasonal effect by a first harmonic describing the effect of the elliptic orbit of the earth round the sun. But what about 1976 when there was a very hot summer and a drought. In order to get a deseasonalised figure do we take out the 'average seasonal factor over the years or do we somehow speculate that we have had a particularly hot summer but that this is no reason to think that the so called deseasonalised level has changed? To me you take your choice dependent upon your use of the resulting figures. But of course in Arthur's case he does not know all the uses to which the figures will be put. To be more concrete on this point, one of our recent clients who is a manufacturer of alcoholic drink has experienced a major rise in demand over 1975 and 1976. This was naturally attributed to great marketing success although what had really happened largely resulted from the unusually hot couple of months in the two years combined with an advantage relative to the other alcohols of no V.A.T. When the summer weather reverted to its more typical pattern and the brewers successfully lobbied for V.A.T. on this particular drink, sales fell dramatically. Clearly in this sort of case it is of vital importance to the Company to estimate how much of its sales is to be thought of as due to relative price advantage, promotions, variation within year etc. But again what is the deseasonalised series? Do we take out the seasonality of advertising, of the earth going round the sun, of the particular weather conditions giving rise to a freak summer, the seasonal buying habits of the customer, which in the case of government institutions may just reflect the current policy of placing orders regularly at quarterly intervals and not any seasonal usage, and so on.

If asked why they like deseasonalised series, many users would remark that they are looking for turning points and trend changes. Perhaps also they are now interested in other forms of change such as jumps in 'underlying level'. For most official series it is probably good enough to filter out regular seasonal effects in the traditional ways, using local linearity and smoothing. However, this is not adequate to deal with sharp changes in the series or in any of its structural components. The absence of any parametric formulation in Professor Dempsters paper surprises me. At this meeting we have heard from Professor Akaike, Adrian Smith and Dr. Makov about such formulations and of course Colin Steven and I have used these successfully for many years under the title of Bayesian Forecasting (1976). The big advantage of the parametric or state-space formulation is that it allows one very easily, to attribute variation to many sources and is, after all, in the spirit of the Bayesian statistical linear model with the associated estimation of effects in terms of their distribution functions.

We have also had success with this Bayesian approach in hierarchical forecasting and gave a paper on this at the Royal Statistical Societies Conference in 1977 (Harrison, Leonard and Gazzard). Here we were commissioned to develop a method of forecasting aggregates and their constituents ensuring compatibility in the sense that the sum of the parts was equal to the whole at all levels. Again with probability distributions this reduces to conditional forecasting and this method has been fully documented by I.C.I. who use the resultant software a great deal.

However in expressing my surprise at the ommission of this type of representation let me admit that I am unfamiliar with Mandelbrot's approach and that since I only got a copy of Arthur's paper yesterday I have had no chance to look up the literature. At first glance I am not attracted to it and on a technicality with equation 3.3 would ask why with $H = 0.5$, the variance $C(0) = O$?

I would close with one important point. It is my personal view that stationarity imprisons us. Surely we must recognise that variables have probabalistic effects. That is if I launch an advertising campaign described only in terms of the amount of money to be spent I am unsure of its effect and hence my view of the future is much more uncertain. I can learn on line about this probabalistic effect and I can model it or I can pretend a naive stationarity and ignore it. The most convenient way of modelling that I know is the parametric approach and I would urge that this be given much more attention.

A. ZELLNER (*University of Chicago*):

The papers by Brown, "Aspects of Multivariate Regression" and by Dempster, "Bayesian Inference in Applied Statistics" are valuable contributions which treat important problems. Since I prefer reverse alphabetical order for an obvious reason, I shall comment on Dempster's paper first and then turn my attention to Brown's paper.

First, Dempster presents a personal view of how Bayesian inference ought to be implemented. According to Dempster, "The objective of Bayesian inference is to quantify uncertain knowledge about a set of unknown quantities in terms of a posterior distribution of those unknowns." (p.1). In connection with this definition, it is important to include as yet unobserved values of variables in the "set of unknown quantities" and to mention predictive distributions. In addition, I believe that Dempster's discussion would be enriched by relating it to various theories of scientific method, for example Jeffreys's and those of other philosophers of science. Without such considerations, it is difficult, if not impossible, to appraise Bayesian and other systems of statistical inference. For example, the issue of whether probability is better regarded as a frequency or non-frequency concept requires analysis in terms of a theory or alternative theories of scientific method.

As regards some specific issues in Dempster's discussion of Bayesian inference, his remarks on "tail area significance tests" should, in my opinion, be expanded to consider Jeffreys's Bayesian significance test procedures. As pointed out in my and Siow's paper for this Conference, for many testing problems Jeffreys-like posterior odds ratios relating to pairs of hypotheses are monotonically increasing functions of tail areas associated with usual t and F statistics used by non-Bayesians in testing hypotheses. Thus there is a direct link between Jeffreys-like posterior odds ratios and tail areas, a relationship which may explain why many applied statisticians have persisted in their use of tail areas or "p-values" in appraising hypotheses.

I am sympathetic to Dempster's view that "Bayesian inference is logically separable from decision analysis" De Finneti at a 1968 conference at Frascati expressed a similar view and I believe that R.A. Fisher and H. Jeffreys also agree with

Dempster's view. However, I.J. Good and others appear to take the position that "quasi-utilities" are generally employed in inference and that inference may be viewed as partially contained within decision analysis. Clearly there is a need for more work on the axiom systems underlying inference (or learning) and decision (or utility) analysis to help resolve the issue of separability. For example, the relation of Jeffreys's and Savage's axiom systems could be studied to determine whether the separability view is logically tenable.

Second, in connection with Dempster's discussion of the practical problem of modelling seasonal time series, let me draw participants's attention to the recently published volume, Zellner, ed. (1978), in which many of the issues discussed by Dempster are treated at length in contributions by a number of statisticians and econometricians. In the volume, three approaches to the analysis of seasonal time series are distinguished, namely (1) the descriptive, non-modeling approach, (2) the statistical modeling approach and (3) the subject-matter causal modeling approach. Dempster's approach is an ingenious example of a statistical modeling approach that he compares with two other statistical modeling approaches, the stochastic seasonal ARIMA approach associated with Box and Jenkins and an AR approach advocated by Parzen and Hiplel and McLeod. One might add to this list the mixed deterministic-stochastic seasonal models developed and applied by Pierce, (1978). While these statistical modeling approaches can yield useful results, it is my view that they must be augmented by a subject-matter causal modeling approach. Without a good subject-matter understanding of the nature of seasonality and factors which produce changes in seasonal patterns, it is the case that mechanistic, statistical models do not have a firm foundation.

Practically speaking, this means that parameters of such models may in fact be variables and thus the difficult problem of assessing good prior distributions for these "parameters" may be intractable. Further, as Ploser (1978) notes, analysis of economic models makes it highly unlikely that the restrictions on the moving-average polynomial-lag operators needed to produce the Box-Jenkins multiplicative seasonal ARIMA schemes will be satisfied in general. Also, the variation of policy-control variables can introduce non-stationary effects in the time series processes for individual variables. These considerations, and others which could be added, point in the direction of devoting more effort toward understanding the causes of seasonality and producing reasonable, serious subject-matter models which will enhance our scientific understanding of seasonal phenomena, for example changing seasonal patterns and differences in seasonal processes for different variables. Bayesian techniques can be employed to appraise alternative subject-matter models, estimate their parameters, and use them for prediction and policy purposes. It could very well be the case that fractional Gaussian processes will be valuable in the context of subject-matter causal modeling of seasonal time series. My impression is that Dempster appreciates these points and plans to devote more attention to them in future research.

I am in full agreement with Dempster on the importance of computation in Bayesian analyses and the need for good Bayesian computer programs. In connection with our NBER-NSF Seminar on Bayesian Inference, we have established a Computation Committee headed by Joseph B. Kadane. S. James Press, a member of the Computation Committee has written a paper, Press (1980), which provides information about a number of computer programs.

With respect to Brown's paper, he is critical of the usual least squares estimate or diffuse-prior posterior mean for the regression coefficients shown in (2.3) and (2.4) of his paper on grounds that it does not take account of the between regressions covariance matrix, $\Gamma = \{\gamma_{ij}\}$. This is not a reasonable critique since (2.3.) is the posterior mean relative to a particular prior and the maximum likelihood estimate based on the normal and other symmetric distributions for the error terms. The fact that the symmetry of the problem results in the regression coefficient estimates not depending on the nuisance parameters in Γ seems to be a blessing and not a fault. Also, Hill's cogent discussion of near admissibility or restricted admissibility of "usual" estimates, such as $\hat{B} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q)$ in (2.3), indicates circumstances in which $\hat{B}$ can be justified as an estimate —see Hill's paper, cited by Brown, pp. 566-568— Whether these particular circumstances obtain is the crucial issue. If they do not, then it is unreasonable to use $\hat{B}$; if they do, it is reasonable to use $\hat{B}$. As Hill remarks, "Thus stable estimation may justify the use of Lebesgue measure and the estimator Y [here, $\hat{B}$], as approximations... but it is important to be aware that the approximation must be justified separately in each usage and that it cannot hold for all y." (p.568). Hill talks of "approximations" because he believes that some prior information is available. On the other hand, Jeffreys views Lebesgue measure as a canonical prior for representing ignorance regarding the values of the regression coefficients and $\hat{B}$ the appropriate, not approximate, estimate given the assumed state of ignorance. Of course, if more information is available, as assumed by Hill and by Stein, it can be employed and will lead to estimates of $B = (\beta_1, \beta_2, \ldots, \beta_q)$ which are not independent of Γ, assuming, as Brown does, that Γ has a known value. For example, in the case of exact linear restrictions on the elements of B, it is well-known that the maximum likelihood estimate will usually depend on Γ.

Rather than assess a serious prior for the elements of B, a chore which Brown considers too onerous, he opts for use of the exchangeability assumptions described on p. 5. It is important to emphasize that these assumptions are hard to defend in many practical applications. In the Lindley-Smith approach $\beta_i \sim N\ (\mu,\Sigma)$ implies that all q regression coefficient vectors have the same mean μ, hardly a satisfactory assumption in many economic applications. Further, the assumption that $\mu_i \sim N\ (\gamma,\ \sigma^2_\gamma)$ implies that the elements of μ have a common mean, again hardly a tenable assumption in many applications. Also, the zero mean assumption in (3.3) leading to (3.4) is tenuous in many applications. However, Brown notes that the zero mean assumption can be relaxed and also writes, "Of course the appropriateness of priors depends on the application..." (p.6). Thus Brown emphasizes, quite reasonably, that one must assess the appropriateness of prior assumptions and I contend this process is not far different from assessing an appropriate prior distribution for the regression coefficients.

I pointed out some years ago that the restrictiveness of the natural conjugate prior for the elements of B, mentioned in Rothenberg (1963) can be avoided by assessing a general normal prior for the elements of B —see Zellner (1971), pp. 238-240.— If we write $\beta' = (\beta'_1,\ \beta'_2, \ldots, \beta'_q)$, the prior which I suggested is:

$$p(\beta,\Gamma) \propto |\Gamma|^{-(q+1)/2} \exp\{-(\beta-\bar{\beta})' C^{-1} (\beta-\bar{\beta})/2\} \qquad (1)$$

a diffuse prior for the elements of Γ (or Σ in my notation), and a normal prior for the pq elements of β, with prior mean $\bar{\beta}$ and prior covariance matrix C. Using this prior, I derived the following approximate posterior mean, **b**, for β:

$$\begin{aligned} \mathbf{b} &= (C^{-1} + S^{-1} + X'X)^{-1}[C^{-1}\bar{\beta} + (S^{-1} + X'X)\hat{\beta}] \\ &= \bar{\beta} + [I_{pq} - (C^{-1} + S^{-1} + X'X)^{-1}C^{-1}]\,(\hat{\beta}-\bar{\beta}) \end{aligned} \qquad (2)$$

where $\hat{\beta}' = (\hat{\beta}_1', \hat{\beta}_2', \ldots, \hat{\beta}_q')$ and $S = (Y-X\hat{B})'(Y-X\hat{B})/n$.

It is seen that **b** is a matrix-weighted average of the prior mean vector, $\bar{\beta}$, and of the least-squares estimate, $\hat{\beta}$, with their respective precision matrices as weights. The second line of (2) puts **b** in a"shrinkage" form where the shrinkage is toward the prior mean vector $\bar{\beta}$. If it is appropriate to specialize (2), for example by giving C a particular form or by assuming $\bar{\beta} = 0$, it is possible to obtain particular "ridge-like" estimates. The critical issue is whether these particular specializing assumptions are reasonable. If they are not, it is unreasonable to impose them. Also, it should be mentioned that (2) is the mean of an approximate normal posterior distribution for β with covariance matrix $(C^{-1} + S^{-1} + X'X)^{-1}$. With additional effort, a better approximate posterior distribution for β could be obtained.

Two issues arise regarding the prior in (1). First, it would be useful to have an informative prior for the elements of Γ. Ando and Kaufman pointed out that a prior in the inverted Wishart from places strong restrictions on the prior variances and covariances of the elements of Γ. While Lindley and Press have made some progress on the problem of formulating an informative prior for Γ, I do not believe that the problem has been satisfactorily solved. As regards procedures for assessing the normal prior for β in (1), an extension of the approach (Zellner, 1972) which I formulated for assessing normal priors in univariate, multiple regression models is possible. Applying this approach to each regression equation yields the prior mean and covariance matrix for each β_i, $i = 1, 2, \ldots, q$, namely $\bar{\beta}_i$ and C_{ii}, the matrices on the diagonal of $C = \{C_{ij}\}$. The extension of the approach involves the assessment of C_{ij} for $i \neq j$, that is $\mathrm{cov}(\beta_i, \beta_j)$, $i \neq j$*. Brown's paper has stimulated me to consider this problem which I regard as tractable.

In summary, I urge Brown and others who utilize procedures based on exchangeability assumptions or ridge-regression procedures to consider carefully the assumptions underlying their procedures. I believe that careful attention to these assumptions will lead to a serious assessment of prior distributions which is required to avoid introducing erroneous information in analyses.

* Further, the assessment procedure can and should include checks on the assumed normal form of the prior.

REPLY TO THE DISCUSSION

P.J. BROWN *(Imperial College, London):*

I am most grateful for the two invited discussions presented here and the verbal contributions from various participants at the symposium. Professor Zellner has a number of points concerning the first part of my paper. Let me comment generally on these. The motivating force of our work is the prior distribution for the regression coefficients in multivariate regression. It seems important to me to delineate sets of archetypal priors, investigate their implications, choosing between these priors, in a practical situation by means of my prior knowledge for the particular situation together with accumulated knowledge of the implications of divergence of behaviour should the prior be inappropriate. Indeed L.J. Savage (unpublished book, The Subjective Basis of Statistical Practice, 1961, Section 2.15) emphasises the fuzziness of held prior opinions. He states "In practical work, I try to take advantage of whatever common properties of the acceptable probabilities I can discern". Exchangeability is a very important feature, valid in some situations but not in others as emphasised in the paper. Furthermore, results such as the sampling theory results of section 4 enable one to investigate theoretically the performance of a class of estimators resulting, relative to the Bayes estimator which corresponds to vague prior knowledge. Synthesis of Bayesian and sampling theory properties is, I think, important.

I very much appreciate the substantial contribution from Professor Harrison. He has indeed hit the nail on the head in questioning the general relevance of least squares and his models deserve careful consideration. I naturally disagree with the nature of his criticism of 'ridge regression'. I think 'ridge regression' provides a range of possibilities of wide but of course no means universal usefulness. Let me answer some of his points in detail.

I echo Jeff's concern for careful comparison, particularly in the I.C.I. study he mentions. I do not wish to criticise the company that gave us both sustenance for a number of years; rather the summary nature of the conclusions stated here. Both 'ridge regression' and even ordinary least squares are not simple well defined techniques. Their application to a practical problem involves various protocols such as, which variables to include, whether to transform them, etc. Additionally 'ridge regression' demands careful scaling of the explanatory variables guided by prior information.

Also of crucial importance is the estimation of the ridge constant. Many of the methods of estimating this constant mentioned in our paper were not available when the I.C.I. study was concluded in the early 1970's. In the absence of detailed evidence I must therefore be rather sceptical about the study.

As far as the study in our paper is concerned one does need to beware of using the PRED criterion in isolation. Continuing Jeff's nautical metaphor, it is a bit of a red herring. Its virtue of simplicity of calculation masks the fact that it is really the probabilities of winning each seat that is important. These probabilities may be summed to give overall predictions. Integrations necessary for their calculation are performed in election night forecasting for the B.B.C. (Brown and Payne, 1975) but I did not go to the trouble of calculating them in this paper. In their absence goodness of prediction is better reflected by such measures as *SD* which measure the closeness of observed and predicted values. For this small subset of the full 635 constituencies in the

United Kingdom there are many models that do well retrospectively using the criterion PRED. In fact a very simple well established model gives the same performance on PRED as M_1, the most flexible of Jeff's models. This model estimates the percentage change for each party and adds the average of these changes uniformly to all the undeclared constituencies. The calculations for $n = 15$ observations declared are given in Table 4 in the fine detail necessary to assure oneself about what is happening in the data.

TABLE 4

Votes (in thousands) and Percentage Changes of Electorate for each party from February to October 1974:

				%			%			%			%
CONSTITUENCY	Electorate (thousands)	C2	C1	ΔC	S2	S1	ΔS	L2	L1	ΔL	N2	N1	ΔN
KILMNOCK	60	9	14	-8	22	24	-3	3	5	-3	15	8	12
GL CEN	26	2	3	-4	9	9	0	1	1	0	3	2	4
FIFE E.	56	16	21	-9	7	7	0	5	8	-5	13	9	7
DUMFRIES	62	18	22	-7	13	13	0	4	6	-3	13	9	7
G.L. PROVN	55	3	6	-6	21	23	-4	*	*	*	11	7	7
G.L. SHETL	38	4	6	-5	13	14	-3	1	*	*	7	6	3
G.L. SPRING	48	4	7	-6	17	18	-2	1	*	*	9	8	2
G.L. CATH	50	16	18	-4	15	16	-2	1	*	*	6	5	2
EDBR N.	47	13	16	-6	8	9	-2	4	5	-2	8	5	6
G.L. GARSC	55	5	10	-9	20	21	-2	2	*	*	12	9	6
AYR	52	17	22	-10	14	17	-6	3	*	*	7	5	4
EDBR PNT	55	14	18	-7	13	14	-2	4	7	-6	10	5	9
GL KELVIN	43	7	11	-9	12	13	-2	2	*	*	6	6	0
GL MARYHL	52	3	7	-8	20	20	0	1	*	*	10	9	2
ANGUS S	52	15	21	-12	4	6	-2	3	*	*	17	15	4
	Averages			-7			-2			-3			+5

Key: * Denotes to contending candidate for that party

Notice that the percentage changes in proportion of the electorate are reasonably stable. Conservative votes are tending to go down by about seven percent, Labour by about two percent, Liberal by about three percent and Nationalists go up by around five percent. There is some evidence that Glasgow constituencies (listed as GL) move less towards the Nationalists (only a three percent increase).

With these estimated changes four constituencies are wrongly predicted on PRED. GL GOVAN and STIRL FG are wrongly predicted as changing hands from Labour to Nationalist. Perth and EP is wrongly predicted as remaining Conservative and finally ROSS and CRM is wrongly predicted as switching from Conservative to Nationalist. Also, aside from re-estimation of average changes as n increases, since all but the last of these four constituencies fall in the 25 to 45 declaration band they will not cause prediction problems after $n = 45$.

Noting that the constant term is generally left unshrunk (section 5.1) this model is actually our ridge model with a very large ridge constant. With a lower value of ridge constant allowance is made for possible dependence on the other variables and indeed the variable for the incumbant party in February is typically of some usefulness in prediction. Looking back we might have envisaged that a Labour incumbancy would have been more valuable than a Conservative incumbancy but the effect is minimal on criteria other than PRED. Overall let me say however that in the election night forecasting context of predicting 635 constituencies, prior information from a multitude of sources is used and more variables entertained than used here. Of some importance, for example, are variables which define the perceived tactical situation in a constituency, for if ones favourite party stands little chance in your constituency you might decide to vote against the party you don't like by voting for the party that you dislike less. In general local information is available from psephological experts, opinion polls and 'post' polls. For details of Election night forecasting as implemented for the B.B.C. see Brown and Payne (1975). The example in this paper does not claim to reflect the multitude of concerns to be found there. Additional experience in the two 1979 elections (General and Direct Election to European Parliament) will be reported shortly.

A valid point to come out for our present paper is I think that the method of estimation of the ridge constant is critical and notwithstanding the success of our study, when the number of observation is small compared with the number of parameters, methods such as HKB and Sclove do mimic least squares too closely. They are not applying the implied prior information strongly enough in these circumstances. In anticipation, in all election work for the B.B.C. we have used a fixed ridge constant specified from previous experience. Reiterating, ridge regression is not a single universal tool but requires careful molding to available prior information.

Our experience of election forecasting does give us confidence that the concerns of 'robustness' and model adequacy as listed under (i), (ii) and (iii) of Jeff's discussion are not compelling. Indeed it seems that here Jeff himself does not find them compelling since his models M_1, M_2, involve the additive assumption of normal homoscedastic error and constancy of parameters over time. Multiplicative analysis of proportions in the voting context is fairly well established. Hawkes (1969) considers various models accounting for the transitions between parties from one election to another. These models are typically rather unstable with respect to election data and seem not to be much used although Miller (1972) has used a version of ridge regression to aid estimation. Model M_1 is somewhat simpler than those models in that in much the same spirit as the ubiquitous 'swing' (average of party i increase and party j decrease in the share of the total vote (or two party vote) it concentrates on the ij transition without accounting for the effects of other parties.

Both models M_1, M_2 have the air of prior distributions constructed retrospectively so as to perform well on PRED. Both models concentrate on the previous winner (party i) and provide no linkage between θ_{ij} for different i. Thus after fifteen constituencies have declared, all of which had been Conservative or Labour held, no information is available on θ_{ij}, i denoting Liberal or Nationalist. Total reliance is on the prior.

If there is a case for a change of model it is that votes be assumed to be Poisson (before conditioning on the electorate size) and that the log of the votes be linear in a set of variables. The computational burden of such a log-linear approach does not seem to be very necessary over the typical range of election data as evidenced by even the simple calculations of Table 4. Also Model M_1 with additive error is to me a little unnatural. Modified to a multiplicative error all the calculations of our present paper could be applied, if so desired, to changes in log proportion of electorate. Further, if S_{jk} is the ratio of October to February votes for party j in constituency k, the model

$$\log S_{jk} = \beta_{Jk} + e_k$$

where β_{jk} has a linear structure in terms of explanatory variables (which could include i, the previous winning party) naturally leads to an appealing measure of "log-swing' from party j to party i given by

$$\beta_{jk} - \beta_{ik}.$$

However, it is easy to lose the inclination to study such modifications when our existing prediction methods perform as well as evidenced in the recent Direct Elections to the European Parliament. I hope you did not miss the B.B.C. programme 'Decision for Europe' (June 10th 1979) which presented them.

A.P. DEMPSTER (*Harvard University*):

I thank Professors Harrison and Zellner for their wise comments, most of which I take as friendly amendments.

I have, over 25 years, spent much time studying the views of many past and present leading thinkers on inference, feeling close to some such as Fisher and deFinetti and more distant from others such as Jeffreys and Savage. I hope one day to develop a reasoned exposition of my position, including its almost total debt to others. For now, however, I think more is to be gained by using the actuality of experience in applied statistics to inform theories of scientific methods than vice versa. In particular, the use of axioms to buttress a largely transparent logical system seems to me less valuable than extended testing of the consequences of the axioms in practice.

Zellner wishes that I would use more econometric theory and causal modelling, and I certainly hope to do so. I do wonder, however, whether the so-called causal models of macro-econometrics have much to do with the real causal factors which necessarily operate at a very micro level. The problems of inadequate information to specify a realistic causal system are so great that causal interpretations of feasible macromodels may do more harm than good, if taken at all seriously.

I agree with Harrison that seasonal techniques should be documented publicly. They also should be defended rationally, which I think means having their Bayesian origins exposed. A good technique needs to make a rational assumption about how much one hot summer should affect one's judgments about the following summer. If good climatic theories (e.g., about ocean temperatures and currents) are available, they should be used, but in the end there will be a residual dependence on unverifiable prior assumptions. Ideally, several different scenarios should be presented so that the naive user can be warned and the sophisticated user can introduce his own prior beliefs.

REFERENCES IN THE DISCUSSION

BOX, G.E.P. and JENKINS, G.M. (1976). *Time-Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

HAWKES, A.G. (1969). An approach to the analysis of electoral swing. *J. Roy. Statist. Soc. A* **132**, 68-79.

HARRISON, P.J. and STEVENS, C.F. (1976). Bayesian Forecasting (with discussion). *J. Roy. Statist. Soc. B* **38**, 205-247.

HARRISON, P.J., LEONARD, T. and GAZZARD (1977). Multivariate hierarchical forecasting. *Proc. Annual Conf. Roy. Statist. Soc.*, Manchester 1977.

MILLER, W.L. (1972). Measures of electoral change using aggregate data. *J. Roy. Statist. A* **135**, 122-142.

PIERCE, D.A. (1978). Seasonal Adjustment When Both Deterministic and Stochastic Seasonality are Present. In *Seasonal Analysis of Economic Time Series,* (Zellner ed.) 242-269, Washington, D.C.: U.S. Government Printing Office.

PLOSSER, C.I. (1978). A Time Series Analysis of Seasonality in Econometric Models. In *Seasonal Analysis of Economic Time Series,* (Zellner ed.) 365-397. Washington, D.C.: U.S. Government Printing Office.

PRESS, S.J. (1980). Bayesian Computer Programs. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys,* (Zellner ed.) 429-442 , Amsterdam: North Holland Publishing Company.

ROTHENBERG, T. (1973). A Bayesian Analysis of Simultaneous Equation Systems. *Tech. Report.* **6315**, Rotterdam: Econometric Institute.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

— (1972). On assessing informative prior distributions for regression coefficients. *Tech. Rep.*, University of Chicago.

— (1978). (ed.) *Seasonal Analysis of Economic Time Series.* Washington, D.C.: U.S. Government Printing Office.

Trab. Estadist. (1980)
Vol. 31 pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 293-318 Valencia, Spain: University Press

6. Bayesian and non-Bayesian conditional inference

INVITED PAPERS

BARNARD, G.A. (*University of Waterloo*)
Pivotal inference and the Bayesian controversy

DISCUSSANTS

DAWID, A.P. (*The City University*)
DE GROOT, M.H. (*Carnegie-Mellon University*)
DICKEY, J.M. (*University College of Wales*)
GOOD, I.J. (*Virginia Politechnic and State University*)
HILL, B. (*University of Michigan*)
KADANE, J.B. (*Carnegie Mellon University*)
LEONARD, T. (*Warwick University*)
LINDLEY, D.V. (*University College London*)
ZELLNER, A. (*Chicago University*)

REPLY TO THE DISCUSSION

Pivotal inference and the Bayesian controversy

G.A. BARNARD
University of Waterloo

SUMMARY

The theory of pivotal inference applies when parameters are defined by reference to their effect on observations rather than their effect on distributions. It is shown that pivotal inference embraces both Bayesian and frequentist reasoning.

Keywords: INFERENCE; PIVOTAL; ROBUST; BAYES.

1. PIVOTAL INFERENCE

1. A *pivotal model* of an inference situation arises typically when we have a relatively precise idea of the way in which the parameters are related to the observations, and a less precise idea of just how the observations are distributed. Thus for example, we may have observations x_i $(i=1,2,\ldots,n)$ for which μ and σ, respectively, are location and scale parameters, but we may not be sure as to the precise form of their distribution. Then we know that the

$$p_i = (x_i - \mu)/\sigma \tag{1}$$

have a distribution which does not involve the parameters, but we may not know exactly what this distribution is. If we suppose that the x_i are *nearly* distributed independently, each in a double exponential distribution, we might suppose that the joint density of the p_i could be expressed, sufficiently accurately, in the form

$$\phi_\lambda(p) = (1-\epsilon)(\tfrac{1}{2})^n \exp -\Sigma|p_i| \; + \epsilon(\sqrt{2\pi})^{-n} \exp -\tfrac{1}{2}\, \Sigma(p_i - a_i)^2 \tag{2}$$

for some ϵ between 0 and 10^{-6}, and for some vector a with i^{th} component a_i. This would correspond with an idea that, less often than once in a million times, the observations were from a 'rogue' normal distribution; but it will become apparent that the role of this small mixture of normality is to be viewed rather differently, as indicating perhaps only part of the small uncertainty in the form of the distribution.*

We use λ to denote the pair (ϵ,a) which serves to specify exactly which member of the family (2) applies in a specific case. Although λ would ordinarily be called a parameter, we call it, instead, a *label*, because its logical role in the inference is different from that of the pair (μ,σ). And the term 'nuisance parameter', which might be used instead of label, we wish to reserve for a somewhat different concept.

The term 'pivotal' was introduced by Fisher, to denote a quantity such as Student's t:

$$t = (\bar{x} - \mu)\sqrt{n}/s_x \qquad (3)$$

which is a function of the observations and of the parameters whose distribution does not involve the parameters. We use the term in the same sense.

2. The elements of a pivotal model of an inference situation are five in number: $\{S,\Omega,p,P,D,\}$. S is the usual sample space, of possible observations and Ω is the usual parameter space, of possible parameter values. p is a mapping from S x Ω to P, the pivotal space. p is called the *basic pivotal*. We suppose that measures are given on S and on P, and that for each θ in Ω the inverse mapping $p^{-1}(.,\theta):P \rightarrow S$ is 1-1 and measurable. D is a set of probability distributions on P, specified by density functions ϕ_λ. It is convenient, though not logically necessary, to assume the distributions in D to be absolutely continuous with respect to each other.

3. For any specified label λ the pivotal model defines a likelihood model L_λ, consisting of the usual triplet $\{S,\Omega,\psi_\lambda\}$ of sample space, parameter space, and probability function ψ_λ

$$\psi_\lambda(x,\theta) = \phi_\lambda(p(x,\theta)).\partial p\ (x,\theta)/\partial x. \qquad (4)$$

In accordance with our usage, a function $F(x,\theta)$ will be pivotal in L_λ iff its distribution, derived from ψ_λ, does not involve θ.

* if there were such a thing as a 'fuzzy distribution', this would convey the idea better.

Now if $F(x,\theta) = G\ (p(x,\theta))$, for some function G, then it is evident that F will be pivotal in L_λ *for every* λ. F will then be called a *robust pivotal*-- defined as a function of observations and parameters which is pivotal in L_λ for every λ.

4. We now introduce the concept of a *separating family* of distributions. The family D is said to be *separating* iff the only robust pivotals are functions of the basic pivotal-- i.e. iff $F(x,\theta)$ pivotal in L_λ for every λ implies that there exists a G such that $F(x,\theta) = G\ (p(\times\ ,\theta))$.

In the pivotal model for which $S = R^n$, $\Omega = R^1 \times R^+$, $P = R^n$, and the i^{th} component of p is p_i in (1) above, we use Lebesgue measure, and D is the family given by ϕ_λ in (2) above, the family D is separating. The steps in proving this are:

(i) If D is complete (in the sense of Lehmann) it is separating. (I owe this remark to Barndorff-Nielsen.)

(ii) The family of spherical normal densities with arbitrary centre is complete.

(iii) If D' is complete, and ϕ is arbitrary, then for any $\delta > 0$,

$$D = \{\phi_\lambda{:}(\exists \varepsilon)\phi_\lambda = (1\text{-}\varepsilon)\phi + \varepsilon\phi_\lambda',\ 0 \leq\ \varepsilon > \delta,\ \phi_\lambda' \text{ in } D'\}$$

is complete.

We may also note the obvious

(iv) If D is separating in a given pivotal model, and if $D' \supset D$, then in the pivotal model in which D' replaces D, D' is separating. All this implies that a very small element of uncertainty in the form of the distribution of the basic pivotal is enough to ensure that the family of distributions is separating.

From now on we assume that the family D is separating.

5. The basic inferential steps which justify the term 'pivotal inference' are of two kinds: (i) Making 1-1 transformations which amount to no more than renaming the entities involved; (ii) conditioning steps. These latter make use of what I have called 'Modus ponens probabilitatis' (MPP), by analogy with Modus ponens of classical logic:

Modus ponens	**Modus ponens probabilitatis**
We know 'A implies B'.	We know* $Pr(B$ given A$) = q$.
We know 'A' is true.	We know that A is true.
Therefore 'B' is true.	Therefore $Pr(B) = q$.

* or 'agree' - see Sec. 10 bellow.

The general procedure of pivotal inference thus consists in transforming the basic pivotal p, 1-1, to another pivotal q which splits into two parts:

$$q = \begin{bmatrix} q_1 \\ \\ q_2 \end{bmatrix}$$

The second part, q_2, is *ancillary*, that is, it is constant on the parameter space, so that its value is known when the observations are known. Then the original pivotal model can be replaced by one for which the basic pivotal is q_1, endowed with the conditional distribution which it has, given the observed value of q_2.

The role of the concept of 'separating family' can now be seen. It is to guarantee that there is an essentially *unique maximally informative ancillary* (MIA). For any two functions f,g, we say that f is more informative than (mit) g iff there exists h such that $g = h(f)$ i.e. $g = h \circ f$ in Bourbaki notation; that is, if the value of g can be calculated when the value of f is known, but not necessarily conversely. If f mit g and g mit f then f and g convey the same information and are regarded as equivalent. The relation 'mit' is a partial ordering on the set of functions of the basic pivotal; and if $f(p)$ and $g(p)$ are both ancillary, the vector-valued function $\binom{f\{p\}}{g(p\}}$ is also ancillary, and it 'mit' each for f and g. It follows that the maximally informative ancillary is unique up to equivalence.

6. The 'conclusion' of a pivotal inference is then a statement of the conditional distribution of the pivotal q_1, together with a statement of the values of the functions of the observations which enter into q_1. From this statement, if desired, a confidence level of, say, 95% can be chosen, and corresponding confidence sets for the parameters can be found; but such an 'arbitrary' choice of confidence level (and 'arbitrary' choice of e.g. 'shortest', or 'one-sided', for the form of the confidence set) means that information is lost at this stage. Thus, it is suggested that the conclusion should be expressed in the form of the conditional distribution of q_1, with the necessary functions of the observations, allowing each reader of the conclusion to form confidence sets in accordance with his specific interests.

7. To illustrate, we consider the case of the example of section 1, where the parameters are location and scale. Here we transform to

$$q_1 = \begin{pmatrix} \bar{p} \\ s_p \end{pmatrix} \qquad \text{(where } ^{-} \text{ denotes, as usual, mean and } s^2_{()} \text{ denotes variance)}$$

q_2 with ith component

$$q_{2i} = (p_i - \bar{p})/s_p. \tag{5}$$

Then since, from (1)

$$\bar{p} = (\bar{x} - \mu)/\sigma, \; s_p = s_x/\sigma \tag{6}$$

and

$$(p_i - \bar{p})/s_p = (x_i - \bar{x})/s_x \tag{7}$$

it easily follows that q_2 is the maximal ancillary. The Jacobian of the transformation from p to q is

$$J = (n(n\text{-}1)/|q_{2,n}\text{-}q_{2,n-1}|)s_p^{n-2} \tag{8}$$

if the last two components of q_2 are regarded as functions of the first n-2 components. Thus, the joint density of the transformed basic pivotals is

$$J\phi_\lambda\,((\bar{p} + q_{21}s_p,\ldots,\bar{p} + q_{2n}s_p)) \tag{9}$$

and if the observed values of the ancillaries are $c_1, c_2,\ldots,c_n$,

$$c_i = (x_i\text{-}\bar{x})/s_x \tag{10}$$

the conditional density of $q_1 = \begin{pmatrix} \bar{p} \\ \\ s_p \end{pmatrix}$ is

$$K(.)s_p^{n-2}\phi_\lambda((\bar{p} + s_pc_1,\ldots,\bar{p} + s_pc_n)) \tag{11}$$

where $K(.)$ here, as later, denotes a normalising constant whose value is determined by the condition that the integral of the whole expression, over the whole range of the variables $\bar{p},s_p$, should come to 1.

If now C is a set in the space of $(\bar{p},s_p)$, such that the integral of (11) over the set C is 0.95, we have

$$Pr\;((\bar{p},s_p) \in C\,|\,q_2 = c) = 0.95$$

and so, by the usual argument, if we assert that in our case $(\bar{p},s_p) \in C$, i.e. that for our observed $\bar{x},x_x$,

$$((\bar{x}-\mu)/\sigma, s_x/\sigma) \in C$$

we have a joint 95% confidence set for (μ,σ), having the usual coverage frequency property.

7. To express the conclusion of our inference in a convenient and easily understood form, without destroying its full informativeness and uniqueness, I propose we should revert to the practice still common in the physical sciences of expressing our information about a parameter in terms of a 'preferred value' and a 'standard error', for example:

$$\mu = x_0 \overset{+}{-} b \qquad (12)$$

which, strictly interpreted, means that our knowledge of μ is equivalent to knowing that $(x - \mu)/b$ is distributed in a standard normal distribution, and that the observed value of x is x_0. A natural extension of this notation to the example we have been considering would be:

$$\begin{aligned} \mu &= \bar{x}_0 \,\text{-.}\, \sigma\bar{p} \\ & \qquad\qquad \psi(\bar{p},s_p) \qquad (13) \\ ln\sigma &= lns_{x0} \,\text{-.}\, ln\, s_p \end{aligned}$$

to be interpreted as meaning that $\bar{p} = (\bar{x} - \mu)/\sigma$ and $s_p = s_x/\sigma$ have the joint distribution $\psi(\bar{p},s_p)$, and that the observed value of $\bar{\bar{x}}$ is $\bar{x}_0$ and the observed value of s_x is s_{x0}. The sign '-.' is intended to *suggest* subtraction (thought what precedes '-.' is a number, and what comes after is a random variable). However, such a mode of expression suffers from the disadvantage that there can be a wide variety of densities ψ, whose properties may be by no means easy to discern from their analytical expression. It seems reasonable, in cases such as the example we are considering, to relocate the distribution so that its mode is at the origin, and then to make a linear transformation of the pivotals if necessary, to secure that in the neighbourhood of the mode the density can be treated as approximately that of two independent standard normal deviates. This means that the second derivatives of the logarithm of the density ψ, taken at the mode, should be unity for the repeated derivatives and zero for the cross derivative. If this is done, the 'preferred values' would be the maximum likelihood estimates of the parameters, and the matrix multiplying the pivotal vector would be the inverse of the information matrix. This would lead to a 'justification' of the method of maximum likelihood in its wider context (i.e. as it is used in situations other than those to which pivotal inference applies), as an approximation, in a certain sense, to an exact pivotal inference. It is important, however, to realise that maximum likelihood estimates here have a direct justification, as those points in the parameter space which will be contained in *any* shortest confidence sets, quite separate from the justification for the use of maximum likelihood in more general cases.

When, as with the example we have been considering, one of the parameters appears as a factor in the error of estimate of the other, special issues arise into which we do not enter in this summary account. This is where we need the term 'nuisance parameter', reserved in section 1 above.

8. In the example we have been considering, we can find a pair of functions of the basic pivotal one of which contains the location parameter and not the scale parameter, while the other pivotal contains the scale parameter and not the location parameter: If

$$t = \bar{p}\,\sqrt{n}/s_p = (\bar{x}-\mu)\,\sqrt{n}/s_x, \qquad s_p = s_x/\sigma \qquad (14)$$

the Jacobian of the transformation is

$$\partial(\bar{p},s_p)/\partial(t,s_p) = s_p/\sqrt{n}$$

and the joint density (conditional on c) of t,s_p is

$$\psi(t,s_p) = K(.)s_p{}^{n-1}\phi_\lambda(s_p((t/\sqrt{n}) + c_1),\ldots,s_p((t/\sqrt{n})+c_n)). \qquad (15)$$

We can now take the marginal density for t by integrating out s_p (after substituting $u = s_pt$, $s_p = u.t$, $ds_p = du/t$)

$$\zeta(t|c) = \frac{K(.)}{t^n}\int_0^\infty u^{n-1}\phi_\lambda\{u((1/\sqrt{n}) + (c_1/t)),\ldots,u((1/\sqrt{n}) + (c_n/t))\}du \quad (16)$$

(showing that under wide regularity conditions on ϕ_λ the tails of the t density behave like $K/t/^n$).

The step of integrating out s_p is an *information-losing* step. Even if we are really interested only in μ, the use only of the marginal distribution of t means that any external information we may have concerning the value of σ and which could give information about the error in μ, becomes unusable. In fact, if we knew, for example, that σ was distributed with density $\Pi(\sigma)$, we should take the integral of (15) after weighting by $\Pi(\sigma)$. While if we knew that, say, $\sigma = 2$, to a sufficient approximation, we should take the distribution of t *conditional* on $s_p = s_x/2$.

9. The possibility that we have, or may acquire, information which enables us to assign a density to σ will be taken into account in the general theory by noting that if σ is assumed to have a known (prior) density $\Pi(\sigma)$ then σ satisfies the definition of a pivotal and should be included in the basic pivotal, which thus becomes (p,σ), with density

$$\phi_\lambda(p\,|\,\sigma)\Pi(\sigma). \tag{17}$$

The maximal ancillary is now larger than before. We can transform from (p,σ) to (d,σ,q_2,s_x), with q_2 defined as in (5) above, and

$$d = \bar{p}\sigma = \bar{x} - \mu,\ \sigma = \sigma, \text{ and } s_x = s_p\sigma. \tag{18}$$

The new maximal ancillary is (q_2,s_x). Making the 1-1 transformation, and conditioning on the observed values c for q_2 and s_x for s_x we obtain, for the joint conditional density of d and σ:

$$\psi(d,\sigma\,|\,c,s_x) = K(.)(1/\sigma)^{n-1}\phi_\lambda\ ((d+s_xc_1)/\sigma),\ldots,((d+s_xc_n)/\sigma))\ldots \tag{19}$$

With this additional information about σ we can improve our confidence statements about μ by basing them upon the marginal distribution of d derived from (19) by integrating out σ. Alternatively, if it is σ we are interested in, we can integrate out d, and obtain a 'quasi-posterior' density for σ which can serve to derive confidence limits for σ if required. This 'quasi-posterior' will be identical with the 'posterior' for σ which would be obtained from the 'improper' uniform prior for μ, independent of σ.

Finally, of course, we may assume a known prior density for both μ and σ, so that the basic pivotal becomes (p,μ,σ). The maximal ancillary will then be the whole set of sample values, or equivalently $\bar{x}$, s_x and q_2, and our conditional distribution will be for (μ,σ), given the sample. It will clearly be identical with the posterior distribution derived in accordance with the usual Bayesian rules.

10. The fact that pivotal inference, as formulated here, *includes*, without *requiring* the use of the standard form of Bayes' theorem is important from the point of view of the Bayesian controversy. The present writer goes a very long way with de Finetti's arguments concerning the way we should react to uncertainty as individuals; as a follower of Wittgenstein I lay less stress on the mental material dichotomy than de Finetti seems to do, but my disagreements here come at a philosophical level remote from applications in statistical or decision making practice. What does differentiate me from many of those who call themselves Bayesian is a respect in which I agree with de Finetti when he stresses the distinction between what he calls the Bayesian standpoint, on the one hand, and Bayesian techniques, on the other. By the latter, which he condemns along with other 'ad hockery', he means the formal applications of Bayes theorem to a prior distribution chosen, not because it corresponds to any individual's actual prior beliefs, but because it has some convenient mathematical property, such as 'smoothness' or 'conjugacy' An essential part of the true Bayesian standpoint is the careful investigation of the prior beliefs of the individual concerned, in the expectation that these prior beliefs will turn out to be peculiar to the individual in question.

If it is accepted that the personalistic Bayesian standpoint is concerned with the coherent development of attitudes in a *single* individual, the question arises as to what function the *statistician* has in relation to his *client* or *clients* where at least *two* individuals are involved. It seems to me that it could be argued, by one who accepts the personalistic view, that the statistician has two functions: (i) he has experience of types of random behaviour-- such as, for example, the likely shapes of measurement error distributions to be found in given circumstances-- which enable him to advise his clients about distributional shapes, and thereby effectively communicate additional *empirical* data, (ii) he then should base his reasoning on those probabilities which can be taken as *agreed* by all parties likely to be involved. Such agreement about probabilities may, in a given case, extend to the 'full Bayesian' case, in which (to refer to our example) the basic pivotal is taken as (p,μ,σ); but in another case there may well be room for individuals to differ concerning their assessment of the prior distribution for μ, in which case the agreed probabilities would extend only as far as the joint distribution of (p,σ). And in yet another case agreement may extend only to the approximate specification of the density of p. In each case the pivotal inference procedure of conditioning on known quantities having known (agreed) distributions can be carried through and the result stated in the form suggested in section 7 above, leaving it to individuals, if necessary, to assess, to within sufficient accuracy (which often will not need to be great) their personal priors with which the statement of the statistical inference should be combined.

To sum up this section, we can say that pivotal inference by-passes the Bayesian controversy by making the inference depend on what is *agreed* between individuals as its basis; how far this goes in the direction of a fully Bayesian inference will depend, in a given case, on how much agreement there is among those concerned. There remains, of course, disagreement with those 'ultra-Bayesians' for whom statistics is a branch of psychiatry, concerned only with purely personal coherence, and who consequently insist that there is no need to ask whether or not there is agreement about assigned probabilities;

and there is also disagreement with the 'ultra empiricists', for whom there is no such thing as statistical 'inference', only 'inductive behaviour'. The rule of Modus Ponens Probabilitatis has as much right as its older, narrower correlative to be regarded as a 'principle' of 'inference'.

APPENDIX

1. We give here the details of the proof outlined in Section 4.

(i) **Theorem:** If $D = \{\phi_\lambda\}$ is complete, D is separating.

Proof: Suppose $F(x,\theta)$ is a robust pivotal, then the mean value of F

$$= \int_P F(p^{-1}(u,\theta),\theta)\, \phi_\lambda(u)\, du$$

does not depend on θ. Hence for any fixed $\theta_0 \epsilon \Omega$,

$$\int_D \{F(p^{-1}(u,\theta) - F(p^{-1}(u,\theta_0),\theta_0)\}\, \phi_\lambda(u)\, du$$

vanishes for all λ. Hence, by completeness,

$$F(p^{-1}(u,\theta),\theta) - F(p^{-1}(u,\theta_0),\theta_0)$$

vanishes for all u. Thus identically

$$F(x,\theta) = F(p^{-1}(u,\theta),\theta) = F(p^{-1}(u,\theta_0),\theta_0) = G(u).$$

(ii) If $\int g(u)\,(\sqrt{2\pi})^{-n} \exp -\tfrac{1}{2}(u-\alpha)'(u-\alpha).du = 0$, all u,

and $g^x(t)$ is the Fourier transform of $g(u)$, then

$g^x(t).e^{it'\alpha-t't/2} = 0$, all t

so $g(t) = 0$ all t

so $g(u) = 0$ all u.

(iii) If $D' = \{\phi_\alpha\}$ and is complete, if $\lambda = \left[\begin{smallmatrix}\epsilon\\ \alpha\end{smallmatrix}\right]$ and

$\phi_\lambda = (1-\epsilon)\phi_0 + \epsilon\phi_\alpha$ for $0 \leq \epsilon \leq \delta$ and if $\int g(u)\phi_\lambda(u)\, du = 0$

for all λ then for all ϵ in $(0,\delta)$ and all α,

$$(1-\epsilon)\int_P g(u)\phi_0(u)\, du + \epsilon \int_P g(u)\phi_\alpha(u)\, du = 0$$

so that

$$\int_P g(u)\phi_\alpha(u)\, du = 0, \text{ for all } \alpha$$

which, by completeness of $\{\phi_\alpha\}$ implies $g(u) = 0$.

2. ON THE BAYESIAN - ANTIBAYESIAN CONTROVERSY

1. It would be foolish to imagine that in the course of what must necessarily be a short paper one could hope to review any more than a few aspects of the issues in a debate which has already gone on for upwards of a century and a half. But of late the controversy seems to have become sharper, with extremists on one side seeming to say that the Bayesian model is the only one which can be used to represent experimental logic, and on the other seeming to say that it should never be used. One is concerned lest such sharp divisions should cause us to lose the respect of the community of experimental scientists which we have only relatively recently gained. It seemed worthwhile to take the opportunity presented by this conference to test whether we are ready to move towards the middle ground.

2. The central aim of the theory of statistical inference I take to be the modelling of the logical structure of experiments with a view to assisting in their interpretation and combination for the advancement of knowledge. In pursuing this aim it has set up many types of logical model, some of which are:

(i) The Significance Test Model (ST model).

Here the elements of the model are the sample space $S = [x]$ of possible experimental results, a 'null hypothesis' H_0 specifying $f_0(x)$, the probability of x if H_0 is true, and a discrepancy function $D(x)$ such that large values of D are thought of as explicable if some alternative to H_0 is true. We calculate the P value, $P = \text{Prob}[D(x) \geq D(x_0){:}H_0]$ and if this is small we are disposed to give serious consideration to the alternatives to H_0. (Here x_0 is the observed result).

The canonical case for this mode of reasoning is provided by Daniel Bernoulli. Asked to consider why the points on the unit sphere representing the poles of the planetary orbits should lie so close together, and why they do not exactly coincide, he began by testing the significance of the departure from a random (uniform) distribution on the sphere. Here $D(x)$ was a measure of clustering, such as the reciprocal of the radius of the smallest circle containing all the points.

It is of the essence of the situation that Bernoulli did this *before* seriously considering alternatives. And to apply Bayes' Theorem he would have had to have given serious consideration to these alternatives.

(ii) The Bayes Model (B model).

Here the elements are S as before, $\Omega = [\theta]$, the parameter space, $f(x,\theta)$ specifying the probability of x if θ is the true value of the parameter, and $Pr(\theta)$, the prior distribution of θ. We calculate the conditional distribution of θ, given x_0:

$$Pr(\theta{:}x_0) = f(x_0,\theta)Pr(\theta)/Pr(x_0)$$

and the posterior distribution represents our conclusion.

The inferential step here consists in conditioning on knowing the observed value x_0, the probability of which is completely specified by the model. It should be noted* that if, in addition to the given four elements we also have a discrepancy measure D, we can calculate a P value as in the ST model and if this is small we may be led to modify our B model. The calculation of the posterior belongs to what George Box has called 'model analysis' and the calculation of a P value belongs to what he has called 'model criticism'.

(iii) The Likelihood Model (L model).

Here the elements are as in the B model, except that $Pr(\theta)$ is missing. If special interest attaches to a particular θ_0, and we have a discrepancy measure D_0 associated with this value, then we can again calculate a P value. If, on the other hand, all values of θ are to be considered on an equal footing, and there are no other logical relevant features in the situation, the inference is given in terms of likelihood, the likelihood function being $f(x_0,\theta)$. For any pair of values θ, θ', the ratio $f(x_0,\theta)/f(x_0,\theta')$ measures the relative plausibility of θ as against θ', on the given data.

A principal disadvantage of the L model is that we cannot, in general, derive the plausibility of a disjunction of hypotheses represented by a range of values of θ. This is because, in general, a disjunction of hypotheses does not specify the probability function of x, nor does there in general exist a function $y = y(x)$ whose distribution is specified by the disjunction. Sometimes such reductions are possible. Thus in the case of a sample from a normal distribution with unknown mean μ and unknown standard deviation σ, the disjunction of hypotheses given by $\mu = \delta\sigma$, $0 < \sigma \leq \infty$, for each δ specifies the distribution of $t = \bar{x}\sqrt{n}/s_x$.

Finally, an L model *may* serve to generate a confidence distribution.

* as pointed out by Box

(iv) The Pivotal Model (P model)

Here the elements are S and Ω as before, together with a space $P = [u]$ of values of a *basic pivotal function* $p(x,\theta) = u$. The fifth element is a family $F = [\alpha]$ of densities on P representing the range of uncertainty we often are in concerning the form of the distribution of the observations x. The parameter α, indexing the members of the family F, is a *model adjustment parameter*. (MA parameter). It is required that for each θ in Ω the mapping $p(.,\theta){:}S \to P$ is invertible, with inverse $p_\theta^{-1}(u) = x$. For each α the distribution of u specified by α yields a probability function of x, depending on θ, given by

$$f_\alpha(x,\theta) = \alpha(p(x,\theta)).\ J(x,\theta)$$

where J is the Jacobian of the invertible transformation $u = p(x,\theta)$, for each θ.

A pivotal model is appropriate typically when we take observations to be normally distributed, when we usually mean that we *think* they are *approximately* normally distributed. Because of the uncertainty in the form of the distribution we can give a precise definition of the parameter θ only by reference to the way in which it affects the observations x rather than by the way it enters the distribution of x.

From a given pivotal model P, for each α we can derive an L model $L(P,\alpha)$, with elements S, Ω, $f\alpha$. In this L model we can define a *pivotal*, following Fisher, as a function of x and θ whose distribution does not depend on θ. In the P model we require, for a pivotal, that it should be pivotal in the L model sense *for every* α. To emphasize this we sometimes call such a function a robust pivotal. It can be shown that, under very weak conditions on the family F, for a function $q(x,\theta)$ to be a (robust) pivotal it is necessary and sufficient that it should be a function of the basic pivotal $p(x,\theta)$:

$$q(x,\theta) = r(p(x,\theta)).$$

If q is constant on Ω it is called an ancillary, and if it is constant on S it is called a Bayesian pivotal. If a function $\phi(\theta)$ exists such that $q(x,\theta) = q(x,\phi)$, and such that for each x the mapping $q(x,.)$ from $\phi(S)$ to $r(P)$ is invertible, then q is said to be a confidence pivotal for ϕ. It can be used to generate a confidence distribution for ϕ.

The inference procedure consists in transforming $p(x,\theta)$ 1-1 to $q(x,\theta)$ where

$$q(x,\theta) = \begin{pmatrix} q_1(\phi_1(\theta)) \\ q_2(x,\theta) \\ q_3(x) \end{pmatrix} \quad \begin{array}{l} \text{is Bayesian} \\ \\ \text{is ancillary.} \end{array}$$

Then when the observations are known, $q_3(x)$ is known, and as in Bayes' argument we can condition on this known value to obtain the joint distribution of q_1 and q_2. The former will give a marginal distribution which yields the posterior distribution of ϕ_1, while the latter will often be a confidence pivotal for a function ϕ_2 of θ, and the mapping from θ to (ϕ_1,ϕ_2) will be invertible.

A noteworthy feature of the pivotal model for inference is, that is always unique, in that the maximal ancillary on which we should condition is unique.

An example of pivotal inference is sketched in the Appendix.

The scheme of pivotal inference can be extended to cover cases where the observations consists of classifications of items into categories; but this involves considerable complication and loss of some of the desirable properties of the model, which is best suited to quantitative observations, discrete or continuous. Over this field it can be seen to cover both the model B and the L model. If the basic pivotal contains a Bayesian component for all the parameters involved, then the maximal ancillary will consist of all the observations x, and the inference will be the usual Bayesian posterior; if the basic pivotal contains no Bayesian component, and if F contains only one element, then we obtain a likelihood model. In general we obtain a mixed model.

It is far from my intention to suggest that the four models listed above exhaust the possibilities. For example the 'predictive sample re-use' models of Seymour Geisser have not been mentioned. Our selection has been made with a view to raising some questions which I hope those present will see fit to answer.

3. The questions are these:

(i) Was Daniel Bernoulli right or wrong to argue as he did? Am I wrong in thinking he could not have used Bayes' Theorem? If so, how would he have used the theorem?

(ii) If it be admitted that the personal theory of probability would always provide complete Bayesian pivotals in the P model, are there not instances where a bevy of Bayesians (in Dawid's useful phrase) might agree on parts of the basic pivotal only, so that the inference could not, with agreement, be carried through to a complete Bayesian conclusion? If so, could not the partial analysis be useful in that it might show that remaining differences of opinion are likely to be unimportant?

(iii) Carrying this situation envisaged in (ii) further, could it not happen that the bevy could agree only on the consituents of an L model? If so, how should they proceed?

It should be clear how I would hope these questions will be answered. If they are so, I think it would be worth emphasis that out differences amount to much less than might be thought.

APPENDIX

Pivotal Inference

Example: S $=$ R^n, Ω $=$ R^1 x R$^+$, p $=$ R^n, $p(x,\theta)$ has i^{th} component $p_i = (x_i\text{-}\theta_1)/\theta_2$, $F = \{\phi_a:\phi_a(u) = \prod_i K \exp\text{-} |u_i|^a + \epsilon, 1 \leq a \leq \infty\}$. Here K, as later, is a normalising constant (not all K's are equal!), ϵ is a small 'error' term expressing uncertainty in ϕ_a sufficient to ensure the 'separating' property -- i.e. that any robust pivotal must be a function of $p(x,\theta)$.

Here the maximal ancillary may be taken as c, with i^{th} component defined by

$$p_i = s_p((t_p/\sqrt{n}) + c_i, \sum_i c_i = 0, \sum_i c_i^2 = n\text{-}1. (i = 1,2,\ldots,n)$$

The Jacobian is of the form $J(c)s_p^{n-1}$ and ignoring the error term the joint density is

$$KJ(c)s_p^{n-1} \exp \text{-}s_p^a \sum_i \,|((t_p/\sqrt{n}) + c_i|^a$$

and in terms of the observations and parameters the transformed pivotals are

$$t_p = (\bar{x}\text{-}\theta_1)\sqrt{n}/s_x,\ s_p = s_x/\theta_2,\ c_i = (x_i\text{-}\bar{x})s_x$$

exhibiting the fact that the c_i are ancillary.

For the complete inference we condition on the observed $c = c_0$, obtaining the joint density

$$Ks_p^{n-1} \exp \text{-}s_p^a \sum_i |(t/\sqrt{n}) + c_{i0}|^a$$

from which joint confidence sets can be obtained. But if we are interested only in θ_1, and ignore the possibility of further information about θ_2, then we can integrate out s_p and obtain the marginal density of t_p as

$$K/ \{\sum_i |(t_p/\sqrt{n}) + c_{i0}|^a\}^{n/a} \tag{4}$$

and we may note that in the case of normality, with $a = 2$, the side conditions on the c_i make this density independent of c_{i0}, and in fact equal to Student's t density on n-1 degrees of freedom. The fact that the condition 1 density in this case does not involve the c_{i0} corresponds to the fact that when the observations are normally distributed $\bar{x}$ and s_x are jointly sufficient for θ_1 and θ_2.

If we find a set T such that the density (4) integrated over T is equal to 0.95, then if $\bar{x}_0$ and s_{x0} are the observed sample mean and standard deviation, the set $\{\theta_1: t_p \in T\}$ is a 95% confidence set for θ_1. The smallest such set will be obtained if T consist of all points t_p for which the density (4) exceeds some suitably chosen constant.

Box and Tiao have discussed this model from the Bayesian point of view, using a 'non-informative prior' for θ_1 and θ_2. For given a, the posterior distribution they arrive at is the same as the confidence distribution derived from (4). That this is not accidental can be seen if we change our pivotal model so that P becomes $R^{n+1} \times R^{+}$, and define the first n components of p as before, but add $p_{n+1} = \theta_1, p_{n+2} = \theta_2$, and regard the $\phi_a(u)$ as giving the density of $p_1,\dots,p_n$, given p_{n+1} and p_{n+2}, and giving to these last two components the distribution corresponding to the prior used by Box and Tiao. In so far as strict Bayesians sometimes object to these improper priors, it might be said that the Pivotal analysis given above is more Bayesian than the Bayesian treatment!.

Box and Tiao also assign a prior distribution to a, on the basis of external information to the effect that the observations are nearly normally distributed, though they are careful to examine whether, over the plausible range of the MA parameter* a, the value of makes any drastic difference. This is, of course, a perfectly reasonable way of dealing with an MA parameter, provided the inferences are suitably qualified. As a matter of fact, for the Darwin data examined by Box and Tiao, it appears more probable, from a reading of Darwin's own detailed account of how he obtained his data, that two of his observations have been given the wrong sign, and that the corrected observations are quite closely normal. If, of course, information was available providing an observational basis for a prior for either or both of θ_1 and θ_2 the pivotal analysis could be carried through on this basis.

AKNOWLEDGEMENT

Research supported by the NSERC of Canada.

* MA stands for "Model Adjustment", MA parameter seems a better term, I think, than "label", or "discrepancy parameter".

DISCUSSION

A.P. DAWID (*The City University, London*):

I have learned to be wary of those who claim that they would like to reconcile the various opposing views on statistical inference. In my experience, the invariable consequence is, rather, a polarisation of attitudes and a great deal of fruitless apoplexy, and Professor Barnard's paper has succeeded in bringing such a reaction. If there is any common attitude that all statisticians might take to Bernoulli's reasoning, it should be that it does not fall within the ambit of *any* of the standard patterns of inference. For example, use of the P-value in the ST model presupposes that the measure of discrepancy is chosen before looking at the data. But if it had happened, say, that the poles of the planetary orbits lay approximately in one plane, rather than being almost coincident, Bernoulli would surely have used a different discrepancy measure, and it seems impossible to correct for this selection effect. This, to me, discredits the ST interpretation of Bernoulli's argument. I do not believe that Bernoulli's reasoning was unsound —it has obvious common-sense appeal— but it is a weakness of *all* modern statistical orthodoxies that they cannot really justify such reasoning.

I don't think it matters much whether or not we can bring about close agreement between proponents of different basic viewpoints. What is important, I believe, is that we should be willing to learn from the insights of our colleagues (both statistical and substantive) of all complexions, and not interpret whatever view we hold so narrowly that we dismiss those insights out of hand. I am happy that this attitude of give-and-take seems to be becoming more common in the statistical community. One area in which I believe it has been fruitful is that of improved estimation in linear models, following Stein's discovery of the inadmissibility of the usual estimator: Hoerl and Kennard (1970), Lindley and Smith (1972), Efron and Morris (1973). More generally, I think that Bayesian ideas will prove extremely valuable to sampling theory statisticians when they come to consider more carefully the modelling process: for example, a Bayesian approach to finite population sampling can be used to justify a superpopulation model (Ericson, 1969). ,This, to some degree, answers Barnard's questions (ii) and (iii), since such a model would represent the agreed component for a bevy of Bayesians who all shared the view that the elements of the population were exchangeable.

M.H. DEGROOT (*Carnegie-Mellon University*):

In the story about Daniel Bernoulli we have an example of the serious difficulty of trying to make inferences about some particular hypotheses from some data when the hypotheses themselves have been suggested by the data. The null hypothesis of a random distribution on the sphere is tested only after it is noted that the poles of the planetary orbits seem to lie close together. A discrepancy function is chosen after the data have been observed, and then evaluated at these same data points. Under these conditions how are we to interpret the calculated P values?

Every set of data exhibits some peculiarities. It would be very surprising if we could not look over some data and then set up a hypothesis H_0 and a discrepancy function $D\ (x)$ that would yield a very small P value based on the same data. But does that mean that H_0 has been discredited? To some extent, perhaps, but not nearly as much as if H_0 and $D\ (x)$ had been selected *before* the data had been observed. How much must we discount the observed significance because of this double use of the data?

The Bayesian approach suffers from the same dangers. We open the newspaper in the morning and read some data on a topic we had not previously thought about. In order to process the data, we try to think about what our prior distribution would have been before we saw the data so we can calculate our posterior distribution. But we are too late. Who can say what our prior distribution would have been before we saw the data. We lost our virginity when we read the paper.

J.M. DICKEY (*University College of Wales*):

I have three points to make on this thoughtful paper by Professor Barnard.

1. The significance test seems to be a more primitive method than the Bayes factor. Thinking is not free, and thinking about alternatives to the hypothesis under test is often more difficult than thinking up a discrepancy measure, or test statistic, to use. There often seems to be an underlying relation between the discrepancy measure and interesting alternative hypotheses, even when it cannot easily be traced.

2. The arguments I have heard made against significance tests seem to be based either on the misuse of tests or on grounds of ideology. ("Whatever is not overtly Bayesian is useless").

In practice, there seem to be two kinds of hypotheses tested: (a) null hypotheses (no-effect models); and (b) working models subject to diagnostic checks. Sample size considerations can play havoc with tail-area tests in the context of (b); less so in (a). A small tail area should not be relied on as an excuse to consider alternatives to a null hypothesis. But if the tail area is *not* small, one is well advised not to bother to build up elaborate theories to explain an apparent effect, for it could very well have been an accident under the null model. (Practical cases where this latter use appear unreasonable tend to involve a poor choice of test, for example, one in which prior information on the variance is ignored).

This limited use for significance tests in context (a) is justified by the inequality,

$$\begin{aligned} B(H) &\geq P\ (T|H)/P\ (T|H^c) \\ &\geq P\ (T|H), \end{aligned}$$

where $B\ (H)$ is the Bayes factor in favour of the null hypothesis H based on the test statistic t, and T is the tail event $\{T \geq t\}$. See Dickey (1977) and references cited therein, and also Good (1950, footnote p. 94). Note that the Bayes factor is approximately the same as the posterior probability for H when it is small and the prior probability is moderate,

$$\begin{aligned} B(H) &= \{P\ (H|D\ )/[1 - p\ (H|D\ )]\ \}/\{P\ (H)/[1 - P\ (H)]\} \\ &\doteq \{P\ (H|D\ )/1\}/1. \end{aligned}$$

3. It is not generally true that the Bayes factor is a monotonic function of the tail area. In the case of a point hypothesis, H: $\mu = \mu_0$ versus H^c : $\mu \neq \mu_0$, write the Bayes factor in terms of the likelihood function $\ell(\mu)$,

$$B(H) = \ell(\mu_0)/\int \ell(\mu)\, p\ (\mu|H^c)\ d\mu.$$

Suppose, as is commonly the case, that the tail area decreases to zero as the maximum likelihood estimate $\hat{\mu}$ goes to infinity. If the likelihood has a location form,

$$\ell(\mu) = f\ (\hat{\mu} - \mu),$$

then it is quite clear that the limiting behaviour of the Bayes factor depends crucially on the relative tail behaviours of f and $p\ (\mu|H^c)$. For example, if the prior density is supported on a bounded set and the likelihood has a tail like a Student-t density, then the Bayes factor will go to unity (no evidence) instead of zero. Data $\hat{\mu}$ very far away, then, will not distinguish between H and H^c (Dickey, 1977). (It might be said to indicate that neither model is reasonable).

I.J. GOOD (*Virginia Politechnic and State University*):

I would like to answer Professor Barnard's question concerning Daniel Bernoulli's use of a tail-area probability in an astronomical context. But I have already given a detailed discussion of tail-area probabilities from a Bayesian or rather "Doogian" point of view in Good (1950, pp. 93-94; and 1976a, pp. 162-165). I would be grateful if people interested in this topic would read those few pages. Perhaps the Editor would regard this contribution as too long if I included copies of those pages here. Some slight impression of the nature of those pages may be gleaned from the following footnote from page 94 of Good (1950):

> "There are two independent reason why the factor in favour of H exceeds $P\ (\chi_0^2)$. The first is that to pretend that the result is $\chi \geq \chi_0$ when it is really $\chi = \chi_0$ is unfair to H. The second is that $P\ (\chi \geq \chi_0|\bar{H}) < 1$, so that the factor from the evidence "$\chi \geq \chi_0$" is
>
> $$P\ (\chi \geq \chi_0|H\ )/P\ (\chi \geq \chi_0|\bar{H}) > P\ (\chi \geq \chi_0|H) = P\ (\chi_0^2).\text{"}$$

After the formal meeting, Professor Barnard drew my attention to Boole (1854, pp.365-368). By using modern terminology Boole's argument can be condensed into the following few lines:

> The final odds of a null hypothesis are equal to the initial odds times the Bayes factor, but we do not usually have pnysycal knowledge of the initial odds nor of the probability of the observed event given the non-null hypothesis.

Although Boole does not (here) mention Bayes, he is in effect saying that Bayes's theorem cannot be used when the appropriate prior probabilities are *unknown* and Boole could therefore be considered to have somewhat anticipated von Mises (1942). They both ignore the possibility of using partially ordered subjective probabilities.

This possibility is not ignored in the 1950 and 1976 references that I have just mentioned. Those references explain why it makes sense to use tail-area probabilities in many circumstances: they often have a loose relationship to approximate Bayes factors. This relationship forms a part of the Bayes/non-Bayes compromise that I advocate and which Professor Barnard should welcome.

It is worth emphasizing that the Bayesian or Doogian explanation of the use of tail-area probabilities shows very clearly how the sample size is relevant: The larger the sample the more the subjective distribution of the statistic, given that the null hypothesis is false, moves away from the distribution given that is the true. Hence a smaller tail-area probability is required to undermine the null hypothesis. For example, if in Barnard's Bernoulli example there had been a million planets, a tail-area probability of say 1/1000 would have been unconvincing for refuting the null hypothesis (that the normals to the planetary orbits were flat-randomly distributed in all directions).

When selecting a significance test criterion we have at least a vague idea of the alternatives to the null hypothesis, and the criterion can be selected as one giving rise to a large expected weight of evidence for distinguishing the non-null hypothesis from the null hypothesis. This weight of evidence (logarithm of the Bayes factor) is based on the test criterion, which does not usually exhaust all the information from the sample.

There are also approximate relationships between Bayes factors based on all the data and tail-area probabilities based on sensible statistics. For example, see Good (1967, 1976b), Good and Crook (1974) and Crook and Good (1980).

B.M. HILL (*University of Michigan*):

Professor Barnard inquires as to the scientific value of Daniel Bernoulli's significance test for the hypothesis of uniformity of the planetary orbits on the celestial sphere. Here we are not considering the various ways in which significance tests are routinely misapplied nowadays by even supposedly well-trained statisticians, but rather the significance test in the hands of a master. Although I must be hesitant to critize a Daniel Bernoulli for anything whatsoever, I would still like to question the value of his tests. The only comprehensible purpose of a significance test without specified alternatives is the purpose of deciding when there is a need to search for new and better models. (In Bernoulli's problem there is in fact a natural alternative, namely coplaner orbits, but Professor Barnard wishes us to ignore this). A *p* value can be used for such a purpose, but so can many other quantities, for example, the surface area of the smallest region of a given shape containing the points, as a percentage of the total surface area. Apart from cases where the *p* value is an approximation to a posterior probability the *p* value has no natural interpretation, and so the question "how small is small" for such a surface area corresponds precisely to the question "how small is small" for a *p* value. Professor Barnard, of course, might not use conventional levels of significance such as .05, .01, but in this case he must tell us how to allow for sample size and choice of the critical region *after seeing the data* in our interpretation of the evidence against the null hypothesis. So I ask what does a *p* value offer over and above simpler and more direct quantitative measures as a guide in the search for better models? Professor Barnard suggests (private conversation) that it allows one to compare different problems on a common scale. However, it seems preferable to me to choose whatever feature strikes one's eye in a particular problem. I see no reason to compare different problems on a common scale. Perhaps Professor Barnard could make clear the purpose of such a comparison. Note that with the approach I am suggesting there would be less likelihood of ascribing statistical significance when there is no practical significance, since it rests upon a more direct perception of the striking features of the data. Often Berkson's interocular traumatic test will suffice.

J.B. KADANE (*Carnegie-Mellon University*):

Professor Barnard rightly calls to our attention the question of the reputation of statistics in experimental disciplines. However I disagree with his diagnosis of the problem: he proposes that sharp divisions among us may lose us respect, while I believe that our reputation lies in the quality of statistics we propose.

Significance testing is a critical point in the philosophical discussions surrounding statistics. The basic question is not so much whether Daniel Bernoulli's use of it was felicitous, but whether we are to endorse present day statistical practice which puts great weight on such tests. Several experiences have led me to conclude that significance testing is much less generally useful than its proponents proclaim. Briefly, some of those experiences are:

(1) (testing a new theory). A distinguished colleague had a new theory (of city sizes) he wished to publish in a statistics journal. The journal insisted on a significance test, so he found the *least* powerful test so that his theory would not be rejected, by the test and by the journal. But he never thought that his theory held *exactly*.

(2) (The catastrophy of too much data). In a sociological study of the frequency of contributions to group discussions, there was a theory Kadane, Lewis and Ramage (1969), wanted to compare to the data. After observing significance at less than 10^{-6}, we found ultimately that plotting the data was much more helpful. This was because we had about 10^4 observations.

(3) (The catastrophy of too little data). A governmental wished to know whether a machine extensively tested in the laboratory worked as well in the field. A significance test revealed "no significant difference", although further analysis showed it was working 75% as well, on the basis of 5 observations costing 1 million dollars each.

In each of these cases enhancing the model and estimating a parameter is much more revealing, although often graphical techniques suffice. There may be an extremely limited role for significance tests, in my view, when the following pertain: (i) the null hypothesis is *honestly* believed by some parties and (ii) the alternatives are expensive to figure out and specify prior distributions for. In such cases a significance test may be understood as a (weak) approximation to a proper Bayesian analysis.

But in my statistical practice, (i) is almost never the case (and (ii) is almost always true!). The only exception for me in recent years is an experiment planned with an astrologer who claimed to be able to distinguish drug offenders from others on the basis of birth dates. Here I put some positive probability on the hypothesis of identical frequency of drug offences. In general, however, the null hypothesis has zero prior probability, and hence zero posterior probability whatever the data. Attempts to rescue even Bayesian versions of hypothesis testing (Dickey (1976) have lead to their abandonment (Kadane and Dickey (1980)).

On Professor Barnard's word that my criteria (i) and (ii) are met in the case of Daniel Bernoulli's application, I do no object to significance testing in this case. But as a general matter, I believe that significance testing threatens the respectability of statistics more than any other single factor.

T. LEONARD (*University of Warwick*):

Professor Barnard has stimulated a general discussion on significance testing on the basis of a practical example with only five observations. Could I simply remark that for larger sample sizes the problem of goodness of fit should no longer be controversial? It is possible to show that we would compare the chisquared statist with the product of the degrees of freedom and the log of the sample size. This approximates the Bayes solution under a very wide range of prior assumptions, and essentially fixes the significance level for any particular sample size. For very large sample sizes it confirms that the standard test is too much ready to reject the null hypothesis.

D.V. LINDLEY (*University College London*):

What ought Daniel Bernoulli to have done? Use a Fisher-von Mises distribution on the sphere and look at the posterior distribution of the spread, particularly in relation to the value of the spread corrsponding to a uniform distribution. (This is effectively what Jaynes described modern physicists as doing, in his discussion of Zellner's and Bernardo's papers). The difficulty with a test of a hypothesis using a tail-area, significance level is that there is always something that is snificant. The introduction of a discrepancy function tacitly introduces the notion of an alternative and hence of the Bayes approach.

A. ZELLNER (*University of Chicago*):

In this interesting contribution, it is indicated that in the *ST* model approach, a discrepancy function $D(x)$ is introduced and no formally stated alternative hypothesis is used. However different choices of the discrepancy function can lead to different results. Could it be that choice of a particular discrepancy function implicitly implies an alternative hypothesis (H_A) which the investigator has in mind?. If so, why not formulate a posterior odds ratio for H_o and the alternative hypothesis, H_A? To be specific, if for a normal mean problem, H_o is the hypothesis that the mean is zero, $\mu = 0$, one might use as a discrepancy function $t^2 = ny^2/s^2$ and compute the *P*-value associated with t^2, i.e., $Pr\ \{t^2 \geq t_0^2 | H_o\}$ where t_o^2 is the observed value of t^2. The problem here lies in the interpretation of the *P*-value. It is not equal to the posterior probability that the mean is zero, as is well-known. Jeffrey's analysis of $H_O{:}\mu = 0$ vs. $H_A{:}\mu \neq 0$ leads to the following posterior odds ratio, $K_{OA} \doteq \sqrt{}\ \pi\nu/2/\ (1+t^2/\nu)^{(\nu-1)/2}$ where ν = n-1, with n the sample size, and involves the 'discrepancy function'' t^2. It is apparent that K_{OA} is a monotonicaly increasing function of the *P*-value and thus, in my opinion, gives a rationalization for the use of *P*-values in this and other problems.

This example illustrates how use of a particular discrepancy function can be rationalized in Bayesian terms. In Bernoulli's problem, with the null hypothesis of a random (uniform) distribution on the unit sphere, it would be interesting to find the alternative hypothesis (or hypotheses) which leads to a posterior odds ratio that is a monotonic function of the particular discrepancy function for the Bernoulli problem mentioned by Barnard and to show how use of various alternative hypotheses affects the form of the discrepancy function. That Bernoulli employed a particular discrepancy function, apparently without justifiying its use should not be interpreted as good statististical practice in general.

In addition, it is the case that Bayes' factor (BF), the ratio of the posterior odds ratio to the prior odds ratio, can be interpreted as an "inverse" discrepancy function. For large sample size in many problems, $\text{-}2\ell n\text{BF} \doteq \chi_q^2 - q\ \ell n \nu$ or $\text{BF} \doteq \nu^{q/2} \exp\{-\chi_q^2/2\}$, where $\text{-}2\ell\ n\text{LR} \doteq \chi_q^2$; with LR = the likelihood ratio, q = the number of restrictions under the null hypothesis, and ν = degrees of freedom. For this large sample approximation, $\chi^2_{\ q}$ can be interpreted as a discrepancy function in Barnard's sense but is not as satisfactory as BF which has a direct interpretation and involves a dependence on χ_q^2 *and* the quantities q and ν.

REPLY TO THE DISCUSSION

G.A. BARNARD (*University of Waterloo*):

Since discussion concentrated on the first part of my paper I will confine my reply to this. I hope the issues raised in the second part may be discussed more fully at another Conference as pleasant and stimulating as this one.

I agree entirely with Joe Kadane and with Morris DeGroot. In their day to day work statisticians are almost always concerned with estimation rather than with hypothesis testing. But the importance of an issue cannot be judged entirely on the basis of its frequency of occurrence. The need for significance tests, such as Daniel Bernouilli's arises at the growing points of science, when a new departure, involving concepts not yet thought of, is required. Such occasions are rare, but their importance cannot be over-estimated. And before undertaking the arduous task of thinking up new concepts we would normally insist on *P* values much lower than the fossilised numbers 0.05, 0.01, or even 0.001; this, at least partly, because we need to make allowance for selection, though the size of this allowance cannot be determined with any precision.

All the other discussants seem to assume that it is just as easy to compute $\Pr(E|\text{ not-}H)$ as it is to compute $\Pr(E|H)$. Only if this is so can we convert the measure of *relative* plausibility given by Bayes Theorem:

$$\Pr(H|E)\ /\ \Pr(H'|E) = (\Pr(E|H)/\Pr(E|H')).\ \Pr(H)/\Pr(H')$$

into an absolute measure by setting H' = not-H. But this is, almost by definition, impossible when not-H involves concepts not yet thought of.

To give just one illustration, in the paper to which Good refers in his contribution, he assumes that it is known that the observations are independent; but such an assumption would often be false in real life. I would suggest that the many and strange forms of dependence that could arise would defeat the possibility of computing $\Pr(E\,|\,\text{not-}H)$ in this case.

In practice we ussually can think of not-H as consisting of the disjunction of a mixture of well specified alternatives (such as Lindley's suggestion, in Daniel Bernoulli's case) with an ill-specified 'something else'. For the well specified alternatives we should quote the likelihood ratio versus H, while for the 'something else' we can have not alternative to the P-value. I look forward to the day when in situations such as those we are considering we will specify, not only H and P, but also a specific (and reasonable) H', with its associated, the likelihood ratio. But we should not pretend to the omniscience involved in assuming that (H or H') exhaust the range of possibilities.

REFERENCES IN THE DISCUSSION

BOOLE, G. (1854). *An Investigation in the Laws of Thought*. New York: Dover.

CROOK, J.F. and GOOD, I.J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part. II. *Ann. Statist.* (in press).

DICKEY, J.B. (1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.* **71**, 680-689.

DICKEY, J.M. (1977). Is the tail area useful as an approximate Bayes factor?. *J. Amer. Statist. Assoc.* **72**, 138-142.

EFRON, B. and MORRIS, C. (1973). Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc. B* **35**, 379-421.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. B* **31**, 195-233.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin, New York: Hafner.

— (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. B* **29**, 399-431. (with discussion). Corrigendum **36** (1974), 109.

— (1976a). The Bayesian influence, or how to sweep subjectivism under the carpet. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. (C.A. Hooker and W. Harper, eds.), Vol. 2, 125-174. Holland: D. Reidel.

— (1976b) On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**, 1159-1189.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

HOERL, A.E. and KENNARD, R.W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics* **12**, 55-67.

KADANE, J.B., LEWIS, G.H. and RAMAGE, J.G. (1969). Horvarth's theory of participation in group discussion. *Sociometry* **32**, 348-361.

KADANE, J.B. and DICKEY, J.M. (1980). Bayesian decision theory and the simplification of models. *Evaluation of Econometric Models*. (J. Kmenta and J. Ramseyu, eds) New York: Academic Press.

LINDLEY, D.V. and SMITH, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. B* **34**, 1-41.

VON MISES, R. (1942). On the correct use of Bayes' formula. *Ann. Math. Statist*, **13**, 156-165.

Trab. Estadist. (1980)
Vol. **31**, pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 319-346 Valencia, Spain: University Press.

7. Personal and inter-personal ethics

INVITED PAPERS

SAVAGE, I.R. (*Yale University*)
On not being rational

KADANE, J.B. (*Carnegie-Mellon University*)
and
SEDRANSK, N. (*State University of New York at Albany*)
Toward a more ethical clinical trial

DISCUSSANTS

LINDLEY, D.V. (*University College London*)
SKENE, A.M. (*University of Nottingham*)
BERNARDO, J.M. (*Universidad de Valencia*)
DEGROOT, M.H. (*Carnegie-Mellon University*)
GOOD, I.J. (*Virginia Polytechnic and State University*)
OHAGAN, A. (*University of Warwick*)

REPLY TO THE DISCUSSION

On not being Rational

I.R. SAVAGE
Yale University

SUMMARY

A Bayesian decision-theoretic approach appears to me as a sensible idealization of a guide to behaviour. At the same time I would like to understand why my behaviour is not always of this form: I sometimes use randomization and I sometimes find confidence intervals acceptable. Not all of my problems have an explicit cost function. Am I lazy or irrational? Do I use non-Bayesian conventions to help communicate? Is the cost of rationality-computation missing from the Bayesian model?

Keywords: DECISION THEORY, UTILITY, COMPUTATION COST, RANDOMIZATION, IMPUTATIONS, CONCEPT FORMATION, NONPARAMETRIC STATISTICS, CONFIDENCE INTERVALS, MAXIMUM LIKELIHOOD, VAGUENESS, CONJUGATE PRIORS, STATISTICAL PACKAGE.

1. INTRODUCTION

Theories never apply perfectly. Serious use of statistical theory quickly runs into major problems. In this essay the emphasis is on difficulties faced by a Bayesian. Other approaches, Waldian or Fisherian, would raise similar problems. Since the decision-theoretic Bayesian model is the most complete of current theories it tends to create the most difficulties when used. I doubt that there is some eclectic statistical theory that can be used without difficulties.

If the range of application of a theory is very narrow then when the theory can be applied the application might be routine. For a short time I had a colleague who held the theory that a statistical analysis could make statements about the data in hand and nothing else. The applications that I have in mind in writing the following are often more complex and less well structured than the standard textbook examples.

The standard theories of statistics are involved in all of the issues to be raised here. In the story about π (Section 2) either the non-Bayesian will suffer the same discomfort as the Bayesian or he will say it is not a statistical issue which shifts the problem to others. The inadequate treatment of cost of thinking (Section 3) appears to be common to all current statistical theories. The tension between the efficient use of standard statistical models and tailor-made procedures is common (Section 4).

Randomization (Section 5) in sampling and assignment of treatments is a very appealing process but it is seldom easy to show the need for it. Current statistical developments raise new topics and some discussion of imputation (Section 6) is appropriate.

Confidence intervals and maximum likelihood estimation (Section 7) have received an extensive non-Bayesian development. In application they often are given a Bayesian interpretation. This raises ethical and educational issues. In Section 8 a few annoying details are mentioned.

Three closely related topics not covered in this essay are data analysis, model making, and concept formation -see Suppes (1966)-. The lack of a mechanism of discovery in the Bayesian framework is crucial in the use of statistics in scientific research. In this conference Box has argued that the Bayesian framework must be inadequate in this respect. Also, the conference papers of Leonard and Dempster are much concerned with this point.

The author does not claim any new results. The references cover the discussion. Perhaps bringing these topics together with a minimum of technical distraction will be helpful. The presentation emphasizes the problems arising from the noninclusion of the cost of rationality in the Bayesian framework. Even if it should be argued that these costs are implicit in theory, it is clear they are not explicit in use.

2. SOME THOUGHTS ABOUT π

At no cost you can win a bottle of sherry if you correctly state the 29th digit of the decimal expansion of π; an incorrect statement yields nothing. In this situation I think I would pick my favourite digit, 7, and expect my chance is .1 to win the sherry. A moments thought tells me I am a bad Bayesian.

What I should do is think about the problem and compute the 29th digit. Since the Bayesian is rational he should be able to perform this task. Even if rational means something more restricted than perfect reasoning, it still must be noted that the usual Bayesian model does not include the cost of computing. So again in this situation the correct Bayesian action is to find the correct answer.

Some reading this story might know the required digit; it has often been computed and it is available in standard sources. In some situations it might be worthwhile to go to the library and look up the digit. If one bottle of sherry is replaced by a lorry load of sherry, I would come up with the correct digit.

Apparently thought is very much like data. One has incentive to do more thinking (more data collecting) when the stakes are increased. Pure thought, stored data, and data not yet acquired are costly ways of removing uncertainty.

I.J. Good (1950, p. 49; 1968, pp. 125 and 129; 1976, pp. 135-136; 1977) has used the terms Type II rationality and dynamic probability in discussing the topics of this section and of Section 3. Also, de Finetti (1975, pp. 278,291) has discused π in this context.

3. THE COST OF THINKING, ANALYZING OR COMPUTING

Recently, Watson and Brown (1978) have discussed the problem of how much value there is in doing an analysis —in the operations research context— *before* the analysis is performed. Their situation must include the analysis required to design a statistical investigation. Watson and Brown's references summarize the related work, including their efforts to find the value of analysis in several case studies.

The costs of analysis do not appear explicitly in statistical theories. In large scale statistical activities, such as a national census, there will be explicit budgeting for items such as planning, data handling, and publication. This process appears to be empirical; theory to help choose optimal amounts of these items is not used. Watson and Brown suggest that empirical evidence would be a good way to solve their operations research problem. It is not clear how well this process of learning can work because of the great variety of complex situations that one needs information about. In the public sector, it often appears that there is inadequate budget for analysis after data are collected. The problem might be that it is relatively difficult to obtain appropriate budgets for soft items like analysis in contrast to hard items, like data.

Economic theory perhaps could make a formal background for the optimal choice of amounts of thought and analysis. Those commodities are known to be valuable but they are hard to value. Most statistical consulting does not have the market mechanism to help establish value.

We will come back to this topic when we discuss standard models, randomization and imputations. To avoid giving the impression that the discussion does not relate to the usual activities of statisticians consider the cost of the following tasks:

(a) Limiting the scope of analysis, determining which variables need analysis.

(b) Specifying joint distribution of all variables, items to be measured and states of nature.

(c) Evaluating losses associated with decisions and states of nature.

(d) Searching for the best kinds and amounts of data to collect.

(e) Determining how much to spend on analysis and communication of results.

If these aspects of the problem are handled properly their costs might be comparable to the usually considered costs, such as sampling costs and terminal losses. It is my impression that we know very little about the correct expenditures on items (a)-(e). At this point statistical theory does not automatically help us to choose good levels for these activities. Subject specialists must be able to help with some of the choices such as limiting scope (a). Careful work on (a)-(e), even if the resulting sample sizes must be reduced, should provide a powerful mechanism to avoid superficial data collection and glib analysis. The difficulty is that we already think we know how to collect and analyse data and it is still challenging, if not frightening, to think about (a)-(e). One is put off from theorizing on these topics because of the unwieldy anticipated results. One fears an infinite regress.

Moore (1978, p. 72) considers some of the above costs are trivial but he incorrectly related them to total costs in contrast to costs due to uncertainty. (He does discuss many problems in applying Bayesian decision theory).

4. STANDARD MODELS

"Assume... are iid..." is an expression seen so often that one is tempted not to check its appropriateness. Although this set of assumptions is used by pushers of nonparametric statistics they are proud of their lack of use of assumptions. Some reflection could lead to different results:

1. The care for experimental detail to assure iid can often yield stronger assumptions such as normality.

2. In fact iid might be replaced by a weaker exchangeability assumption.

Two standard models of great importance are the packaged programs and the Raiffa-Schlaifer conjugate priors. These examples well illustrate the advantages of standardization: A great variety of problems can be handled at low costs. A convenient mode of communication is developed. Many people can use advanced technology. The problem is to make sure that these advantages greatly exceed the disadvantages: One can force a situation into the wrong model. One can be lazy and not take full advantage of the available options. One can unwisely restrict the kinds of data to be obtained just so a standard model can be used.

The ideal is to have the standard models available and that their use is supervised by skilled individuals. The net results are to increase resources for research and to make sure that unusual situations receive appropriate attention.

5. RANDOMIZATION

A Bayesian is about to sample a finite population. Should he take a random sample? If a random sample costs no more than a grab then why not random sample? If the Bayesian acted as if the population elements were exchangeable then random sampling has no disadvantage. (For a discussion see Ericson [1969].) In this situation there is an advantage to the Bayesian in taking the random sample, even if he is unhesitating about the exchangeability. In particular, random sampling gives others confidence that the work has been done properly.

This confidence might be at two levels. The use of random sampling is an indicator that the whole job has been done with professional care. For some, there will be increased acceptability of the results because they feel random sampling is a necessary part of a valid procedure. How much the Bayesian should pay to buy confidence of others is not clear.

When would the Bayesian have an aversion to random sampling in the above situation? This would happen if he did not really accept the exchangeability assumption; he might really prefer some form of stratified sample. In fact, the acceptability of random sampling can be used as a form of self examination of the Bayesian to tell if he strongly believes in exchangeability.

The Bayesian might use randomization as a technique to avoid expensive activities such as thinking. Thus in the current example let us assume his interest centres on the total income of the population. He might well have many variables that he could use for stratification, such as age, sex, address, education, profession, number of children, etc., etc. Regardless of the costs of the various types of samples the Bayesian may conclude it is more economical to ignore the other variables and just work with income. He saves thinking about the complex multivariate distribution of all of the variables. This averted task is one where there is limited experience. He might have other substantial savings in data collection and analysis.

If the Bayesian followed this simple path then he might even be willing to pay for the randomization for his own peace of mind. To put this into a formal analysis could be awkward.

Rubin (1978 a, Sect. 5) contains a technical discussion of some of these comments. His 1974 article is also relevant. Savage (1962, pp. 34, 88-89) vividly describes a Bayesian's problem with randomization.

6. IMPUTATIONS

It is common in handling large data sets to have missing values. Those are replaced by imputed values. The data set is then ready for analysis. No matter how much or what kinds of analysis are to be performed there usually is just one imputation process applied to a set of data.

The great advantage of doing the imputation is that statistical methods for complete data sets are much simpler than those for data sets with missing values. So if there is going to be much analysis it is less expensive to do one costly imputation and many routine procedures than many complex missing data procedures.

If computing costs did not dominate then presumably imputations would not be used. For it is hard to believe that one set of imputations will be correct for a variety of problems involving different loss functions and different prior distributions. Again, the Bayesian probably cannot afford to think it out in detail —thinking is expensive— so that he is inclined to use the single imputation process. (See Rubin —1978b— for a more technical discussion).

7. CONFIDENCE INTERVALS AND MAXIMUM LIKELIHOOD ESTIMATION

The first few times I was told about these procedures it sounded like gibberish. The instructors knew the correct definitions and attempted to present them. The definitions are awkward and appear to be about the wrong thing. Compare:

(a) The *mle* is that value of the parameter which would have maximized the probability of the data.

(b) Given the data the *mle* is the most probable parameter value.

Or:

(a) In the long run the procedure for 95% confidence intervals will create intervals including the parameter 95% of the time. On any particular occasion the probability of coverage is either 0 or 1 but there is not evidence from the data to say which value is correct.

(b) A 95% confidence interval includes the parameter with probability 0.95.

Both of the (b) statements are false. Many users of the procedures have (b) and not (a) statements in mind. In a Bayesian framework the (b) statements are good approximations when the samples are large or the prior is diffuse. Since these are important and commonly used procedures how should the statistical community reduce the large number of errors? I am convinced the non-Bayesian can do nothing. They have had little success in fifty years of expositing; their message is useless. Perhaps the Bayesian should help the users of the (b) statements to understand the implications. This also might not be useful for most people don't want to expand their formal knowledge of statistics. The Bayesian can often let well enough alone.

8. FINE TUNING

In practice it is hard to even begin to be a Bayesian. One generates inconsistent prior distributions. Computation of exact probabilities would be meaningless. Utilities are often not even approximated. Since the theory does not provide for the expense of these costly activities it is not surprising that the behaviour required by the theory does not occur.

Lindley, Tversky and Brown (1979) present a mechanism for the resolution of inconsistencies (they assume no cost for this mechanism). Without giving any details here, it seems appropriate to suggest that in investigating ways to remove inconsistencies one should not discard the possibility of the Bayesian doing further introspection.

Many discussions of axiom systems for the Bayesian have appeared. Suppes (1974) specifically evolves theories which do not require the Bayesian to give exact values. Again, the cost of accuracy is not in the model.

Absence of utility measurement in much of applied Bayesian statistics can reflect a variety of causes, such as lack of interest, excess cost, or unable to produce at any cost. It does seem possible that a statistician could present probabilities that would be moderately acceptable to all interested parties. On the other hand the interested individuals might have widely varying utilities. An example of some interest is the allocation of funds from central to local governments. This is now often done by formulas using social and economic data. The utilities of the civil servant statistician, the executive, the legislative body, the local governments, pressure groups, and the people might all be different and all hard to approximate. Even for such major activities this work is seldom begun. It would be costly and it would be, technically, hard to justify. At this time the evidence regarding the usefulness of such analysis is ambiguous.

"The coherent individual is supposed to assess his probabilities and utilities for everything. Of course, taken literally, this is absurd; but it does not invalidate the theory any more than the failure of the claim to predict the whole future of the universe, given the position and velocities of particles now, invalidates Newton's theory" Lindley's discussion of Suppes (1974, p.181).

REFERENCES

DE FINETTI, B. (1975). *Theory of Probability: A Critical Introductory Treatment*, Vol. 2. London: Wiley.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion) *J. Roy. Statist. Soc. B* **31**, 195-223.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

— (1968). Corroboration, explanation, evolving probability, simplicity, and a sharpened razor. *British J. Philos. Sci.* **19**, 123-143.

— (1976). The Bayesian influence, or how to sweep subjectivism under the carpet. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol. 2 (C.A. Hooker and W. Harper, eds.), 125-174. Dordrecht, Holland: D. Reidel.

— (1977). Dynamic probability, computer chess, and the measurement of knowledge. In *Machine Intelligence 8* (E.W. Elcock and D. Michie eds.), 139-150. New York: Wiley.

LINDLEY, D.V., TVERSKY, A. and BROWN, R.V. (1979). On the reconciliation of probability assessments (with discussion). *J. Roy. Statist. Soc. A* **142**, 146-180.

MOORE, P.G. (1978). The mythical threat of Bayesianism. *Int. Statist. Rev.* **46**, 67-73.

RUBIN, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educat. Psycol.* **66**, 688-701.

— (1978a). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6**, 34-58.

— (1978b). Multiple imputations in sample surveys a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Section of the Amer. Statist. Assoc.*, 183-188.

SAVAGE, L.J. *et. al.* (1962). *The Foundations of Statistical Inference — a Discussion.* London: Methuen.

SUPPES, P. (1966). Concept formation and Bayesian decisions. In *Aspects of Inductive Logic* (Hintikka J. and Suppes P. eds.). Amsterdam: North-Holland.

— (1974). The measurement of belief (with discussion). *J. Roy. Statist. Soc. B* **36**, 160-191.

WATSON, S.R. and BROWN, R.V., (1978). The valuation of decision analysis. *J. Roy. Statist. Soc. A* **141**, 69-78.

Toward a more ethical clinical Trial

J.B. KADANE and N. SEDRANSK
Carnegie-Mellon University
and
State University of New York at Albany

SUMMARY

Current methods of conducting clinical trials require the patient to agree to have his treatment assigned randomly, where his individual characteristics are taken into account only to balance the treatment groups. A Bayesian alternative involves eliciting the prior opinions of the group of clinicians who designed the study. Each patient is then guaranteed that the treatment he will receive is the best for him either in the opinion of at least one individual clinician or as a consensus of several, given the patient's characteristics and all the information available from the trial when the assignment is made.

Keywords: ETHICS, DECISION THEORY, UTILITY, ELICITATION OF OPINION, ACCEPTABLE TREATMENT, RADOMIZATION.

1. INTRODUCTION

In this paper, we explore the notion that every patient in a clinical trial should be assigned a treatment responsibly believed to afford therapeutic advantage to him.

The clinical trials considered here involve several different treatments administered to patients who arrive sequentially. In general, a patient must receive treatment shortly after arrival. Patients may be heterogeneous with respect to features or attributes recognizable prior to the determination of treatment and likely to affect prognosis. A well-defined measure of therapeutic efficacy is assumed.

For example, consider a study of patients admitted to a hospital's trauma unit where two new methods of preventing sepsis are being evaluated in comparison with surveillance. Intensive care is given to every patient in the unit. Patients arrive following massive trauma and require immediate attention. Recognizable characteristics include type of trauma (accident or post-surgical complication; head injury or not; long-bone fracture or not) and type of patient (age; probably general physical condition immediately prior to trauma). Proportion of hospital time free from sepsis is taken to be the measure of efficacy.

Commonly in this type of clinical trial, a patient receives a treatment drawn at random from the set of treatments under study. The probabilities of selection for the treatments are fixed throughout the clinical trial and usually are equal. The randomness may be unconstrained, or an overall study design may be based on random selection of one of the possible permutations of a sequence of treatment assignments including a fixed number of assignments to each treatment. Further structure for the design may be imposed by generating separate sequences of assignments for different types of patients.

The most frequently cited motivation for these randomized designs is removal of the treatment selection from the control of the attending physician. This reduces his ability to manipulate the treatment assignments, and hence decreases the possibility of confounding effects of treatment and prognostic factors in the observed results. The use of constrained randomized designs is promoted in order to increase the efficiency of the study (to reduce the variance or mean squared-error of treatment effect comparisons). Also, hypothesis tests about treatment effects can be based on the permutation distribution induced.

Ethics for such a trial have been justified by arguments like the one given by Gilbert, McPeek and Mosteller (1977):

> "Let us consider the question of whether a present patient should give up something for future patients. We, or our insurance carriers, pay the monetary cost of our care. What we do not pay for is the contribution to the medical system by past patients. These patients, through their suffering and participation in studies, have contributed through their illness and treatments to the present state of evidence for all patients. Such contributions cannot be purchased by money but can be repaid in part by making, when appropriate, a contribution to the same system. One good way is through participation in well-designed clinical trials when the patient falls into the limbo of medical knowledge... Thus the patient has an interest not only in the trial he or she has the opportunity to engage in, but also a stake in a whole system that produces improved results that may well offer benefits in the future, if the patient survives the present difficulty. Thus, the social system will likely offer benefits through the larger system even when a particular component of the system may fail to pay off directly for a patient, his family, friends, or some other social group he belongs to".

Need for such an argument arises when some patients are asked to accept a less efficacious therapy under study in order to treat later patients more knowledgeably. If he enters the trial shortly after it is begun, when differences among the effects of the several treatments may be imperceptible or unknown, the patient's sacrifice, if any, may be slight. However, if he enters the trial after a substancial amount of data is available, treatments which appeared equally likely to prove efficacious at the outset may no longer be equally desirable. In this case the patient's exchange of expected therapeutic benefit for expected information may be markedly to his detriment.

This approach asks the patient to accept whatever therapeutic disadvantage may come his way in the name of scientific progress. Such an emphasis on the greater social good relative to the legitimate interests of the patient is less than satisfactory. In this paper we seek an alternative which incorporates new information as it is generated by the trial to protect patients from inadvisable treatments.

2. MODELING THE CIRCUMSTANCES OF A CLINICAL TRIAL

A clinical trial may be proposed in order to reduce controversy about the relative merits of the therapies to be studied and/or to gain information about the efficacy of one or more of the therapies in the absence of any strong prior opinions. In a study to evaluate several therapies, there may be agreement among responsible scientists, physicians in this case, about some of the treatments, disagreements about some and lack of firm opinion about others.

In general, it is reasonable to assume that a set of prevailing opinions within the scientific medical community is represented by various physicians involved in the clinical trial. Establishing and defending criteria for selecting the "prevailing opinions" to use is outside the purview of this paper. It is usefull to express each of these opinions about the efficacies of the therapies studied as a probability distribution for the efficacy measure conditional upon the prognostic factors considered important by one or more of the physicians involved. Once "prevailing opinions" are expressed as distributions, the acquisition of data during the conduct of the study permits updating in the usual fashion. Thus at each point during the study, all opinions are "current" an important divergence from a classical, fixed design for a randomized trial.

Two distinct sets of utilities are involved in the conduct of a clinical trial. One, obviously, is defined in terms of accomplishing the study objectives, *i.e.*, reaching a consensus about preferred treatments and/or acquiring information about treatments. The role of this set of utilities is akin to that of efficiency measures or power function requirements for tests of hypotheses in conventional (non-Bayesian) designs for clinical trials. The second set of utilities is the set of patients´ utilities. For each patient, this utility function represents his own valuation of therapeutic results; there is no apparent counterpart in conventional designs for clinical trials.

Much of the difficulty in reconciling ethics with efficiency in the design of clinical trials seems to arise from ignoring the patients' utilities, or from confusing the two sets of utilities with each other, or arguing as Gilbert, *et al.* do that it is reasonable to assume that the sets are the same.

The two sets of utilities govern our experimental design in different ways: the former as an objective function to be maximized, the latter as a constraint on the solution set.

In the formulations to be discussed, each patient's set of utilities is used to define an acceptable set of possible treatments for this particular patient. Then the selection of treatment from this set is made with respect to the overall scientific objectives of the study.

3. IDENTIFYING ACCEPTABLE TREATMENTS

An acceptable treatment for a patient is considered to be one which in some sense maximizes the patient's interests, where these are expressed as a utility function; an unacceptable treatment is one which in no sense maximizes the patient's expected utility. The patient may himself express a utility function or a general form for the utility function may be supplied for him. In either case, the patient himself is not assumed to have a prior opinion, although he has at his disposal the collection of current "prevailing opinions". Thus the expected utility for the patient reflects his own utility function and the expert opinion he consults. Following the custom of "seeking a second medical opinion", there is a set of expected utilities for a particular treatment corresponding to the set of (updated) prevailing opinions.

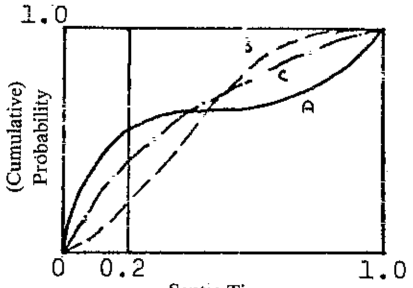
For a study of T treatments where P prevailing opinions are available, a particular patient will have a $T \times P$ array of expected utilities $[u_{tp}]$, where u_{tp} is the expected utility of each treatment t according to the updated opinion of expert p. Define a treatment t as *acceptable* if there is some set of convex weights for prevailing opinions $\{w_p\}$ satisfying $w_p \geq 0$ for all p and $\Sigma\, w_p = 1$, such that

$$\sum_p u_{tp}\ w_p \geq \sum_p u_{t'p} w_p \tag{1}$$

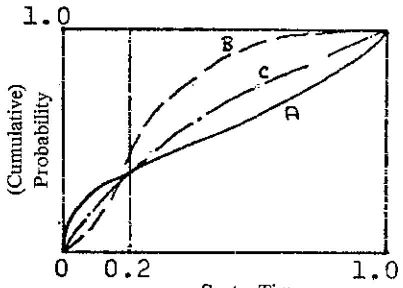
for all other treatments, t'. Acceptable treatments include the treatment most-favored by each expert, and possibly other, generally well-favored treatments. *We propose that each patient be guaranteed an acceptable treatment.*

As an example recall the sepsis-prevention study in the trauma unit. For

the three treatments, A, B, C, two hypothetical prior opinions of participating physicians are shown below for the proportion of hospitalization time likely to be spent in a septic state.



Opinion 1

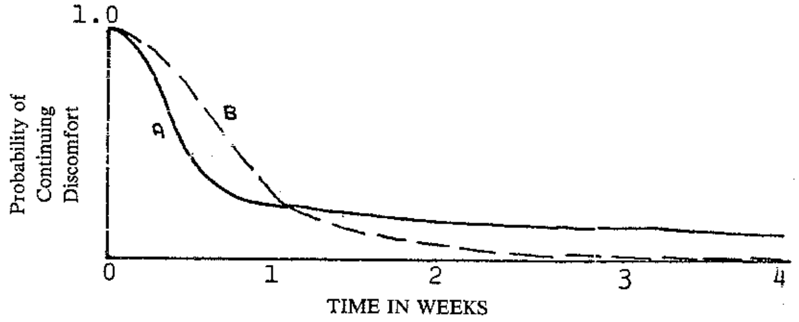


Opinion 2

In this case the patient's utility function may be supplied for him, in view of his inability due to ignorance as well as to incapacitation, to express his own interests. Thus medical knowledge is imposed, in this case by recognition that total septic time in excess of 20 percent of hospitalization time is decisive for a patient's recovery. Hence expected utility is defined here to be the probability that less than 20 percent of hospitalization time is spent a septic state.

Using the prior opinions depicted, all treatments are acceptable, since the three vectors of weights $w_1 = (1,0)$, $w_2 = (0,1)$, $w_3 = (1/2, 1/2)$ satisfy inequality (1) for the three treatments, A, B, C, respectively.

However, even for identical patients the acceptable set of treatments may differ, due to differences in utility function. As an example consider a clinical trial of "soft" contact lenses. In this case, the objective may be the minimization of adjustment time (duration of accelerated eye-fatigue and increased eye-strain). A single expert opinion about the lenght of adjustment time might have the form depicted below.



Thus the expected utilities for two patients who measure utility differently, that is immediate adjustment (0.5 week of discomfort) and eventual adjustment (2 weeks of discomfort), will be maximized by lenses of types A and B, respectively. A third patient whose utility is defined in terms of one week of discomfort will be indifferent between the two types of contact lens. Consequently, the acceptable sets of treatments (lens types) differ for the three patients.

Finally, note that the acceptable set of treatments is defined for each patient, as the patient arrives. Thus the opinions used in the calculation of expected utilities are the original prior distributions (representing "prevailing prestudy opinions") *updated* by all accrued data. Hence, the definition of acceptable treatment is current for each patient at the time his treatment must be determined.

4. DESIGNING WITHIN THE CONSTRAINTS

Restricting treatment selection to the set of acceptable treatments does not, in general, completely specify the design for a clinical trial. Under this restriction the ethical considerations are satisfied for any design, therefore other criteria can be used to determine the design.

For example, a group of physicians committed to the idea of randomized clinical trials could use a random process to choose among acceptable treatments for a patient. In this case, definition of the proper permutation distribution and proper consequent analysis would be complicated greatly. Furthermore, it is logically inconsistent to discard philosophy and to ignore the experts' opinions in this aspect of the design.

Consider, therefore, a criterion based on the overall scientific study objectives, *i.e.,* reaching a consensus and/or acquiring information about the treatments. The relevant set of utilities is defined for the experts in terms of the information to be gained following treatment of the patient. Thus the treatment is selected from the acceptable set to maximize progress of the study, expressed as a function of the experts expected utilities.

A fully optimal sequential design would take into account the history of the clinical trial, including patient characteristics, assignments and results. It would require specification of a probability distribution characteristics of future patients. It would also require specification of a probability distribution for future patients' utility functions in order to consider the acceptable sets of treatments for the future patients. In face of such complexities, we restrict attention to myopic designs, treating each patient as if he were the last one to be studied.

There are two distinct aspects of defining the treatment selection procedure. First, a reasonable utility function must be determined for each expert. Second, and conceptually more difficult, individual utilities must be agregated to form a group decision.

For a single expert, Raiffa and Schlaifer (1961) propose choosing treatment t to maximize

$$\int_X dx \max_d \int_\Theta d\theta \, \mathrm{V}(d,\theta,t,x) \, p(\theta \mid x,t) \, p(x \mid t) \qquad (2)$$

where $\theta \in \Theta$ is the parameter, $x \in \mathrm{X}$ is the outcome of the experiment, V is the expert's utility function, $d \in D$ is the decision reached by the clinical trial, in Lindley's (1971) notation. Here D is the set of possible recommendations of treatments at the conclusion of the clinical trial. Note that D may include decisions of equivalence among a subset of preferred treatments, as well as selection of a single recommended treatment for each patient. Good (1956), Lindley (1956) and Lindley (1971) suggest maximizing expected information over possible experiments, in this case possible treatment selections. Bernardo (1979) shows that maximizing information can be treated as a special case of maximizing expected utility.

In general, expected utilities for experts will differ because of initially differing opinions, whether or not the experts' utility functions have a common form.

Suppose that treatment t belongs to the acceptable set for the current patient. Denote by V_{tp} the expected utility of treatment for the expert holding prevailing opinion p. For the first patient, without loss of generality, $0 \le V_{tp} \le 1$ for all t and p, since $\{V_{tp}\}$ are unique only up to a positive linear transformation (see Savage, 1954) and hence can be standardized with $\max V_{tp} = 1$ and $\min V_{tp} = 0$ for each p. This standardization may be repeated with each subsequent patient.

Alternatively the expected utilities for the first patient to enter the trial can be standardized in the foregoing manner. Then the standardization coefficients for each expert can be used throughout the remainder of the study. In this case the range restriction on V_{tp} will not necessarily hold for patients after the first.

Selection of treatment can then be made to maximize a suitable function of $\{V_{tp}\}$. For example, one treatment selection procedure is given by

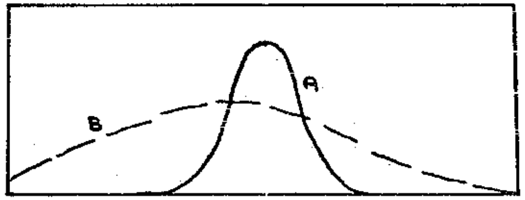
Choose t to maximize $\min_p V_{tp}$.

Clearly, a variety of other measures of aggregate utility are possible, as well.

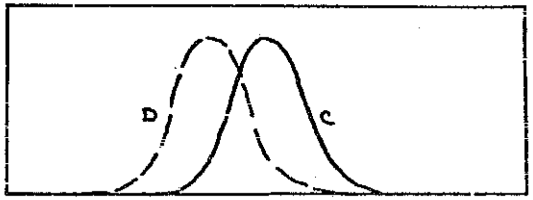
5. CRITIQUE

For clinical trials conducted in an atmosphere of conflicting views the formulation given thus far seems workable. However, studies undertaken in the absence of prior information are vulnerable to premature discontinuation or to unwarranted degeneration of the design. In the definition of the set of acceptable treatments, this difficulty arises from the limitations of determining expected utility as a single number for each treatment. The result is an (unnecessary) narrowing of the definition of acceptable treatment, with resultant loss of flexibility in the overall study design.

Consider two possible situations, one in which there is very little information about one of the two treatments, the other in which there is ample information about both. For a particular expert, the prior opinions for these two cases might be depicted by the densities shown below.



I: PRIOR DENSITIES
TREATMENTS *A* AND *B*



II: PRIOR DENSITIES
TREATMENTS *C* AND *D*

There are smooth monotone utility functions for which the expected utilities calculated for case I and II are the same, with treatments *A* and *C* considered acceptable. In Case II, the rejection of treatment *D* may be considered desirable; whereas in Case I it would be desirable to consider both treatments *A* and *B* acceptable.

It is useful that different patients may express different utility functions resulting, for example, in Treatment *A* being acceptable for some patients, Treatment *B* being acceptable for others, as in Case I. However, one objective of a clinical trial design is that the study be viable without dependence upon a broad distribution of patient utility functions. For the case where diagnosis is

inconclusive, Lindley (1975) describes circumstances permitting valid inference, despite the identical utility functions for all patients. This case is considered further by Good (1978).

Several alternatives deserve investigation. Note that the difficulty arises when the treatment with smaller expected utility also has the more diffuse prior distribution. This suggests that Treatments *A* and *B* may both be considered acceptable because their expected utilities differ by less than some $\epsilon > 0$.

Since a resolution of this problem acts as a governance on the study, adequate solution is essential to the viability of the method.

6. IMPLEMENTING A CLINICAL TRIAL

Many of the essentials for carrying out a clinical trial are available now; others require only moderate efforts to be developed. Representations of "prevailing opinions" must be elicited from physicians holding these views. When then measure of efficacy can be assumed to have a normal distribution or a lognormal distribution, a member of the conjugate prior family can be elicited using the methods of Kadane *et al.* (1978) in the univariate case or of Dawid *et al.* (1979) in the multivariate case.

Updating of prior distributions can be done automatically as data is acquired; and for the conjugate family this can be accomplished quite easily. The major technical difficulty in this regard is the incorporation of censored observations, particularly when the lognormal model is used. Seeking adequate approximations may provide the most effective solution to this problem.

Substantial commitment of programming effort will be required to develop and implement efficient algorithms for the definition of the set of acceptable treatments and for the treatment selection procedure.

Finally, careful selection for a pilot effort should include the following favorable circumstances: primary objective of resolving sharp conflict of opinion (case where the formulation seems to have least vulverability), modest rate of accrual of patients, and single prominent measure of efficacy with clear definition.

7. CONCLUSION

In reply to John Tukey (1977),

> "Many of us are convinced, by what seems to me to be very strong evidence, that the only sources of reliable evidence about the usefulness of almost any sort of therapy or surgical intervention is that obtained from well-planned and carefully conducted randomized, and, where possible, double-blind experiments [see review papers of Byar *et al.* (1977) and Peto *et al.* (1977)]. Dare we prevent outselves

from obtaining reliable evidence?";

the only word we question is "randomized".

8. ACKOWLEDGEMENTS

This research was supported in part by the National Science Foundation under grant SOC-77-27575. The authors are indebted to Morris H. DeGroot and James M. Dickey for helpful discussions before the meeting in Valencia.

REFERENCES

BERNARDO, J.M. (1979). Expected information as expected utility. *Ann. Math. Stat.* **7**, 686-690.

BYAR, D.P., SIMON, R.M., FRIEDEWALD, W.T., SCHLESSELMAN, J.J., DEMETS, D.L., ELLENBERG, J.N., GAIL, M.H. and WARE, J.H. (1976). Randomized clinical trials: Perspectives on some recent ideas. *N. Engl. J. Med.* **295**, 74.

DAWID, A.P., DICKEY, J. and KADANE, J.B. (1979). Matrix *t* and Multivariate *t* Assessment. *Tech. Report*, Carnegie-Mellon University

GILBERT, J.P., MCPEEK, B. and MOSTELLER, F. (1977). Statistics and Ethics in Surgery and Anesthesia. *Science* **198**, 684-689.

GOOD, I.J. (1956). Some terminology and notation in information theory. *Proc. IEEE Part C* **103**, 200-204.

— (1978). Ethical Treatments. *J. Statist. Comput. Simul* **7**, 292-295.

KADANE, J.B., DICKEY, J., WINKLER, R., SMITH, W. and PETERS, S. (1978). Interactive Elicitation for the Normal Linear Model. *Tech. Report*, Carnegie-Mellon University.

LINDLEY, D.V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986-1005.

— (1971). *Bayesian Statistics; A Review*. Philadelphia: SIAM.

— (1975). The effect of ethical design considerations on statistical analysis. *Applied Statist.* **24**, 218-228.

PETO, R., PIKE, M., ARMITAGE, P., BRESLOW, N.E., COX, D.R., HOWARD, S.V., MANTEL N., MACPHERSON, K., PETO, J. and SMITH, P.G. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *B.J. Cancer* **34**, 585; **35**, 1.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard Business School.

SAVAGE, L.J. (1954). *The Foundations of Statistics*. New York: Wiley.

TUKEY, J.W. (1977). Some Thoughts on Clinical Trials especially Problems of Multiplicity. *Science* **198**, 679-684.

DISCUSSION

D.V. LINDLEY *(University College London)*:

Unlike Savage, I feel comfortable with the Bayesian position. I know of no case where it gives an unsatisfactory result and there are many cases where it produces an answer which is more satisfactory than others. This is not to say that there are no difficulties: there are, but they seem to be the sort that should yield to an adequate amount of research effort.

In the case of π, the resolution may lie in removing the excessive formalism sometimes imposed. There is a story that every paper appearing in the Annals of Mathematical Statistics had to have (X, A, P): a triple in which A is a σ-field of events. But why should we have a σ-field? In the case of π the σ-field is complicated and we could not do all the probability assessments demanded of it. What we do is to give *some* p-values, but not all. The important point is that those values given must be coherent. This, I believe, is essentially de Finetti's resolution.

Randomization is a puzzler to a Bayesian. Consider a trial to compare two medical treatments, T_1 and T_2. Suppose the result of the trial is that $p(\mathrm{R}|\mathrm{T}_1) > p(\mathrm{R}|\mathrm{T}_2)$ where R is the event of John's recovery, so that T_1 is preferred to T_2 for John. Suppose however there was a random quantity X such that $p(\mathrm{R}|\mathrm{X},T_1) < p(\mathrm{R}|\mathrm{X},T_2)$ for all X: would T_1 still be preferred for John? Such a set of inequalities constitutes Simpson's paradox. The paradox can be avoided if in the trial, X and the treatments are independent. Now a random allocation of treatments is, by the Bayesian meaning of random, independent of any X, so that a Bayesian might prefer randomization, though independence is what he is really after.

Rubin, (1978) shows that with randomization, the Bayesian calculations are much simpler. Simplicity is connected with the cost of rationality, mentioned by Savage. There has been little investigation of this and Savage does the conference a service by drawing our attention to the problem. When is it worth drawing a decision tree? Clearly the answer must depend on how well the utilities and probabilities can be determined. We usually draw the tree if they can: otherwise we might content ourselves with the initial utility. In the paper by myself, Tversky and Brown that Savage mentions, we studied the assessment of probabilities and suggested using measures of precision associated with them, rather like other determinations in science. Although this made the calculations complicated, and so more costly, my own feeling is that such measures are essential to any resolution of this important problem.

At this moment in statistics, my advice is to try the Bayesian paradigm. I think you will find that if works rather well.

The difficulties with Simpson's paradox also arise in considering the ideas put forward by Kadane and Sedransk. For suppose that the experts were all affected, either consciously or subconsciously, by X; then could it not happen that the resulting confounding of the treatment allocation with X would viciate the conclusions of the trial? Of course, if the confounding were recognized, it might be possible to allow for it: the real danger lies in an unrecognized confounding.

The criterion (1) has an attractive property: namely, it is invariant under linear transformations of the utilities and hence of the expected utilities. For suppose u_{tp} is replaced by $\ell_p u'_{tp} + m_p$, possibly a different transformation for each expert, then the linear form $\Sigma u_{ip} w_p$ becomes $\Sigma u'_{ip}(\ell_p w_p) + \Sigma m_p w_p$ with the final term independent of p and the new weights $\ell_p w_p$ (it is not necessary that these add to one). I feel this is important since utilities are arbitrary up to linear transformations.

The authors admit the possibility that the patient may use his own utility function or have one supplied for him. May he not do the same for the probabilities? It is not clear to me that in evaluating my probabilities, I should use the expert's stated values, for I may feel him to be biased in some way. Thus if the motoring organization tells me their probability of getting stuck in the snow, I shall use a smaller value for my probability, since I believe they exaggerate the hazards in order to discourage people from using the roads and so reducing their chances of having to assist them in snowy conditions. As L.J. Savage pointed out, what the patient really needs is the expert's likelihood function (*not* his probability) to update the patient's prior.

A.M. SKENE *(University of Nottingham)*:

Professors Savage and Kadane though discussing quite different topics are both grappling with problems of utility. The first paper is essentially concerned with the utility structure of the decision problem "How shall I analyse this data?" while the second is concerned with the thorny problem of whose utilities to consider in clinica trials.

Professor Savage is concerned that the practical Bayesian doesn't practice what he preaches and claims that the decision theoretic framework doesn't take the cost of thinking into account. Now I believe I'm rational yet I would also guess the 29th digit of π and hope to win the bottle of sherry. It follows that in taking the decision to guess as opposed to computing the elusive digit I prefer the expected return from guessing viz 0.1 (Sherry - ϵ) + 0.9 (Nothing -ϵ) to the return (Sherry - Effort of computing - ϵ), where ϵ is the small effort necessary to make the snap judgement between the two alternatives. Thus it appears that I think that nine tenths of a bottle of sherry is not worth the extra effort. If I am happy with the outcome of this exercise in self enlightenment then my claim to rationality is unshaken for the time being at least; otherwise I must ponder afresh the fallibility of my decision making process.

There are two issues here. First, there is nothing in the decision theoretic framework which prevents one from including the cost of thinking. It can be incorporated quite naturally (in theory at least) as one of the attributes in a multi-attribute utility function. Secondly there may be some positive advantage in retrospection. It can be very much easier to see what a particular decision implies for the utility function, than attempting to assess the utility function directly. In medical decision making, for example, it is very difficult to get a Physician to assign utilities and costs to various treatments particularly when there is the possibility that a patient may be incorrectly allocated a treatment which is positively harmful. Having attempted such an exercise however, the Physician can then be observed making decisions on a long series of patients and from this information some idea of his actual utility function can be gained. (By considering which misclassification rates are considered aceptable for instance). The value of this exercise lies in the fact that reconciling the two utility functions so obtained may lead to better decisions in the future.

Consider a Statistician, invited to assist in the analysis of a set of data. He sees

several ways in which he might proceed and must choose one strategy. While the factors which influence the utility he has for each strategy are the personal choice of the decision maker, he might, for example, consider how far each strategy satisfied the experimenter's objectives and his own interests, the financial reward involved, the time necessary to execute each strategy, the computing cost/effort involved and perhaps how closely each strategy adhered to the principles of Bayesian statistics. In reaching a decision, the Statistician would, of course, find it necessary to choose weights reflecting the relative importance of these attributes. This situation is surely not unfamiliar. Perhaps we should be asking ourselves what weight we would give to the last of these attributes or, like the Physician, be thinking through the decision in abstract and then observing how we act in practice.

Turning now to the paper by Professor Kadane, I accept the author's remark that the paper is primarily a statement of intent; not a polished work where all the issues have been resolved, but rather an enunciation of a possible direction in which to proceed, together with the problems which are likely to be encountered.

In certain types of societies the practical solution to the ethical clinical trial problem could be achieved quite easily. When the allocation of a treatment depends on whose utility function you consider —the patient's or the physician's— we must combine the utilities in some way.

If the society is such that it sees it as the right or duty to define the role of an individual in a clinical trial then the problem vanishes. In the absence of such political involvement however the concept of an acceptable treatment seems an interesting one and worth investigating, though I am somewhat skeptical that these ideas will, in fact, lead to a new type of clinical trial.

A major problem as the author points out, is that in trials where there are no strong prior opinions it is possible for the procedure to converge to the wrong treatment and this leads the author to the idea of 'nearly acceptable' treatments. However, in practice no patients are denied treatment and in the standard randomised trial patients are all already receiving acceptable or nearly acceptable treatments. In effect, the current argument against randomised trials is based on the premise that 'nearly acceptable' is not 'ethical'.

It may sometimes happen that the patient's utility functions prevent certain treatment comparisons. Consider, for example, a trial comparing mastectomy with a form of radiation therapy for breast cancer. Given that all the participating physicians believe that there is little difference in efficacy and that effective treatment means survival as opposed to death then the utility functions of women involved will reflect preferences between the secondary consequences of the treatments, and thus, for example, the radiation therapy may be universally preferred.

Instances such as this of course don't prove that such a trial will never work. What is of greater concern is the possibility that such a trial is feasible but is misused. Will the utility functions of the participating experts be allowed to reflect things like loyalty to a particular company or the need to justify a particular research project to guarantee future funding? It is just conceivable that under the guise of an 'ethical clinical trial' more patients receive a less efficacious treatment than in a randomised trial. This would certainly be possible in trials where many patients were admitted to a trial before the first results were known. Here, presumably, the experts could continue to use their prior opinions until the first results came to hand.

J.M. BERNARDO *(Universidad de Valencia)*:

Professor Kadane points out that "an emphasis on the greater social good relative to the legitimate interests of the patient is less than satisfactory". I think we must distinguish between the patient's interests *before* and *after* he is known to be affected. For, it seems likely that, before he has got a particular disease, he maximizes his *personal* expected utility by voting a law which will oblige him to accept participate in a clinical trial were he to become ill and the trial necessary.

Indeed, he must balance his better chances of survival because of general scientific progress with the risk of having to accept a particular less efficacious therapy.

M.H. DEGROOT *(Carnegie-Mellon University)*:

Since so much of the discussion of the various papers at this meeting has had a theological tone, it would seem appropriate to introduce the theological terms *probabiliorism* and *probabilism* to help describe the situation considered in this paper. In Webster's Third New International Dictionary we find the following definitions:

Probabiliorism - a theory that in moral questions where certainty is impossible only the more probable course may be followed.

Probabilism - a theory that in moral questions where certainty is impossible any course may be followed that is seen as solidly probable either through clear perception of the principles involved or through awareness of the support of judicious sound authority...any solidly probable course may be followed even though an opposed course is or appears to be more probable.

The authors seem to be urging us to be probabilists in carrying out clinical trials. The patient, however, must strongly hope that his doctor is a probabiliorist.

I.J. GOOD *(Virginia Politechnic and State University)*:

I have often wondered whether most clients who are given confidence intervals use them in some sense as Bayesian estimation intervals; see, for example, Good (1969, p. 184). Perhaps a sample survey is needed to answer the question at any moment in history, and for any field of application.

I have proposed a way of combining judgements of quantiles of distributions by various judges or experts in Good (1979) by methods rather different from those of Lindley, Tversky and Brown (1979). The application that directly provoked my work on this problem was the estimation of mineral resources. This application was brought to my attention by Dr. Larry S. Mayer.

To Professor Kadane, I have to say that, as a patient, I would be not happy with treatment A if two or more clinicians recommend treatment B and only one recommends treatment A, if I had no reason to prefere one clinician's judgment to those of others. I would prefer to accept a majority vote.

Lindley (1975) had an interesting idea for improving the ethics of medical trials, but in Good (1978) his idea was shown not to be as applicable as it at first seemed. I said there that one way to make clinical trials ethical is to *pay* people to undergo them, and I doubt if this proposal was original, in fact it is already done when patients are given free treatment in exchange for entering the trial. Another way to pay patients, if they happen to be prisoners, is to give some remission of sentence as the form of payment. An objection that was raised in conversation by Dr. Kadane is that some are sentenced largely to keep them off the streets. To meet this objection the judge could be allowed to pass such sentences as: "Ten years without the right to enter medical trials and a further twelve years but with the right to enter such trials".

Another idea for making medical trials more ethical, when there is very little to choose between some treatments, would be to arrange to administer the treatments *simultaneously* to a sample of patients, perhaps at numerous medical centers. But this proposal might seldom be practicable.

A. O'HAGAN *(University of Warwick)*:

All statistics in practice is approximate. Perfect analysis requires an infinite amount of effort to achieve. Therefore the ideal of rationality must be tempered by pragmatism. This theme lies at the back of several papers at this meeting but I think Professor Savage hits the nail on the head when he relates it to the cost of effort. The degree of approximation finally accepted in any analysis results from a balance between the gains accruing from more nearly optimal decisions which might be made with an improved approximation, and the cost of that improvement. The cost of better approximation may have many components, but the costs of thought and of computer time spring quickly to mind. Professor Savage suggests that it may be possible to measure these costs, but I doubt if that would help much because the measurement of the gain from improved approximation is much more difficult. It obviously depends on the true analysis, which is unknown, and any attempt to theorise about it will introduce new quantities which themselves must be approximated in practice. Statistics will always be a matter of subjective, unformulated judgements. As Professor Good says in his paper, "I stop when the guessed expected utility of going further becomes negative if the cost is taken into account".

The fact that no practical statistics can ever be more than an approximation to the ideal Bayesian analysis is no reason to despise Bayesian principles and theory -that is the trap into which Dr. Leonard nearly falls with his paper at this meeting. Theory serves at least two distinct purposes. First it provides guidelines. If we know that a certain analysis is optimal for a given problem which we can think of as approximating our own problem, then that analysis serves as an initial approximation for us. Some thought about ways in which the real problem deviates from the theoretical one suggests (by reference to other theory) ways in which we should modify our initital analysis. Dr. Leonard acknowledges this role of theory but gives the impression that it is unimportant, yet without the guidance of theory the applied statistician would be completely lost.

The second purpose is to reduce costs. A new piece of theory means that in appropriate circumstances the statistician can proceed inmediately to a greater accuracy of approximation with only slight costs in extra thought or computing. Professor Savage recognises this, particularly in section 4. The investment that the theoretician's thought represents can yield rich dividends for the practitioner.

REPLY TO THE DISCUSSION

I.R. SAVAGE (*Yale University*):

As noted earlier I claim no originality for this essay. Good's (1979) review of Colodny (1977), hints at my borrowing from L.J. Savage as expressed in his late essay "The Shifting Foundations of Statistics".

This Conference's success should be evaluated in terms of its helping to create theory and application of Bayesian statistics. In doing this there is no last word. I am glad I had the opportunity to participate and I'm thankful for the lively remarks of the discussants.

J.B. KADANE (*Carnegie Mellon University*) and N. SEDRANSK (*S.U.N.Y. at Albany*):

We thank Professors Bernardo, DeGroot, Good, Lindley and Skene for their attention to the problems we pose and for their useful ideas.

The point made by Professor Skene and also raised by Professor Bernardo, that societies differ in the degree of coercion they exert on their members, is unarguable. Hence, the balance between the rights and interests of ill citizens seeking the "best possible" treatment and the rights and interests of the rest of the society in fostering medical research is not uniquely defined for all nations and societies. In the United States, it has been required for some time that a patient consent to participation in a clinical trial prior to the beginning of any therapy or procedure under study, and further that the patient's consent must be given with full knowledge of all relevant information currently available. In this context, both the legality and the feasibility of a clinical trial revolve about the question of whether or not a patient rationally would give informed consent to participate in the clinical trial. Thus the statistical design and analysis must address this question and a Bayesian approach provides a natural formulation. (Consideration of optimal legislation to define a new context for human experimentation is beyond the scope of this paper).

The intent of this paper is to formulate a model for clinical trials which would embody both the rights of the patient and the rights of society and which would exploit the variety of expert opinions within the scientific community to justify study of alternative therapies. Both Professors DeGroot and Lindley express concern that the patient be allowed to modify the expert opinions and/or reject selected expert opinions altogether. When a patient considers entering a clinical trial, he acquires a collection of experts beyond the particular physician he consults directly. A very sophisticated patient might want to correct for the several physicians' varied biased; this modification presents fewer mathematical difficulties than practical obstacles. A much less sophisticated patient might choose to ignore all opinions except that of the physician he consults. However, this model for a clinical trial returns to that of an "uncontrolled" trial, which offers less assurance that the best treatment will be identified correctly.

Sources of potential vulnerability of this class of designs are viewed with concern by both Professors Lindley and Skene; but whereas Professor Lindley considers the possibility of unrecognized factors, Professor Skene worries over infelicitous configurations of known factors. The confounding of unrecognized factors with treatment effects can vitiate the results of any study. To the extent that an unrecognized factor is correlated with a recognized factor, its effect is controlled by the incorporation of the recognized factor in both the statistical design and the statistical analysis. Of course, if the unrecognized factor is independent of all the recognized factors, its occurrence in a pattern resulting in confounding its effect with the treatment effect requires two similar sequences for the sequence of treatment assignments and for the sequence of values for unrecognized factor. (Then purely probabalistic arguments apply; and, for example, in the case of an unrecognized binary factor, the risk of confounding is minimized for balanced designs). Untoward influence of recognized factors, can, as Professor Lindley points out, be averted by proper design and analysis. Precisely for this reason, it has been presumed throughout the paper that covariates are used directly in the probability distributions, and therefore are included in both the design and the analysis.

It is certainly possible, as Professor Lindley suggests, that a physician could inject his (knowing or unwitting) bias into the clinical trial by way of his prior distribution. One strength of the class of designs proposed is its *lack* of vulnerability to a single physician's bias. Only in the event of a rather uniform bias on the part of all experts should the confounding occur, a somewhat less likely possibility than the occurrence of a significant bias on the part of a single physician.

Professor Skene's concerns about failures of the clinical trial design caused by known factors, pose much smaller problems since these possibilities can be examined specifically for each trial before it starts. The possibility that the patients' utility functions might prevent certain treatment comparisons is not well illustrated; in fact, in the circumstances Professor Skene cites, no patient rationally would agree to a randomized trial. Professor Skene also expresses apprehension that less than honest scientists could exploit this class of designs. The use of fraudulent experts can ruin a clinical trial of almost any design; in the class of designs proposed here, as in all designs for scientifically responsible research, experts with conflicts of interests are assumed to be ineligible for influential decision-making roles. The possibility that there is insufficient agreement among experts or that there is near unanimity preventing either initiation or termination of a clinical trial can best be examined by simulation; comment must be deferred until these simulations have been completed.

Professor Good's notion that a trial might be made ethical by simultaneous administration of treatment at several medical centers bears some resemblance to what is now done. But doesn't it address the ethical issue by failing to generate relevant data for any of the patients? A more sequential approach would yield early returns and might avoid giving bad treatments to at least some of the patients. Payment for prisoners in direct form or by sentence reduction is specifically prohibited within the U.S.; and remuneration to non-prision participants may not be of such a magnitude as to impair the individual's judgment of the medical merits of the options offered him.

All these issues are difficult, and well worth discussion and further research. We are grateful to our discussants for their stimulating thoughts.

REFERENCES IN THE DISCUSSION

COLODNY, R.G. (ed.) (1977). *Logic, Laws and Life: Some philosophical considerations.* Pittsburg: University Press.

GOOD, I.J. (1969). What is the use of a distribution?. In *Multivariate Analysis-II* (P.R. Krishnaiah, ed.) 183-203. New York: Academic Press.

— (1978). Ethical treatments. In *J. Statist. Comput. and Simulation* 7, 292-295.

— (1979). On the combination of judgements concerning quantiles of a distribution with potential application to the estimation of mineral resources. *J. Statist. Comput. Simulation* **9**, 77-79.

— (1979). Review of Colodny (1977). *J. Amer. Statist. Assoc.* **74**, 501-502.

LINDLEY, D.V. (1975). The effect of ethical design considerations on statistical analysis. *Applied Statistics* **24**, 218-228.

RUBIN, D.B. (1978). Bayesian Inference for causal effects: the role of randomization. *Ann. Statist.* **6**, 34-58.

Trab. Estadist. (1980)
Vol. **31**, pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 347-382 Valencia, Spain: University Press.

8. Sensitivity to models

INVITED PAPERS

FREEMAN, P.R. (*University of Leicester*)
On the number of outliers in data from a linear model

BOX, G.E.P. (*University of Wisconsin*)
Sampling inference, Bayes' inference, and robustness in the advancement of learning

DISCUSSANTS

EDDY, W.F. (*Carnegie-Mellon University*)
O'HAGAN, A. (*University of Warwick*)
BERNARDO, J.M. (*Universidad de Valencia*)
BROWN, P.J. (*Imperial College, London*)
DAWID, A.P. (*The City University*)
DICKEY, J.M. (*University College, Wales*)
GOOD, I.J. (*Virginia Polytechnic and State University*)
SMITH, A.F.M. (*University of Nottingham*)

REPLY TO THE DISCUSSION

On the number of outliers in data from a linear model

P.R. FREEMAN
University of Leicester

SUMMARY

This paper reviews models for the occurrence of outliers in data from the linear model. The Bayesian analyses are all closely similar in form, but differ in the way they treat suspected outliers. The models are compared on Darwin's data and one of them is used on data from a 2^5 factorial experiment.

The question of how many outliers are present involves comparison of models with different numbers of parameters. A solution using proper priors on all parameters is given. On two trial datasets it is found to be insensitive to choice of priors on all except the parameters representing the amount of contamination in the outliers. Here, choice of even a slightly "wrong" prior can be very misleading. Moreover, it is difficult to choose an appropriate prior when contaminations can be both positive or negative.

Keywords: CONTAMINATION; LINEAR MODEL; MODEL DISCRIMINATION; OUTLIERS; PROPER PRIORS; SPURIOUS OBSERVATIONS.

1. A VARIETY OF MODELS

Consider the common problem in which a statistician would like to use a standard linear model to represent the generation of a dataset arising from some experiment. He has, however, some doubts about whether all the observations were generated by that model and feels there is a chance that some (hopefully, a few) observations will have been contaminated in some way. Recording errors, temporary changes in experimental conditions or the use of abnormal experimental units are the kinds of flaws he has in mind. In analysing the data he must therefore elaborate his simple linear model in some parsimonious way so as to guard himself against such gremlins and ensure inferences about the parameters of interest that are robust.

The word "outlier" will here be used to mean any observation that has not been generated by the mechanism that generated the majority of observations in the datatest. Note that we automatically assume that outliers are a small minority of the observations and that for each possible alternative we must use a different model for outlier generation.

In this section we shall briefly review three such models. We first establish some common notation.

We write the standard linear model as

$$y = \chi\beta + e \tag{1.1}$$

where y is n x 1 and χ is n x p.

If a particular subset $y_{i1} \dots y_{ir}$ of the y's are suspected of being outliers, we partition the y vector into $y_{(r)}$ and $y_{(n-r)}$.

A simple application of Bayes theorem shows that β, the parameter of interest, has posterior distribution that can be written

$$p(\beta\,|\,y) = \Sigma w_{(r)} p_{(r)}(\beta\,|\,y)$$

where the summation extends over all 2^n possible partitions of y, $w_{(r)}$ denotes the posterior probability that the subset $y_{i1} \dots y_{ir}$ are indeed outliers and $p_{(r)}(\beta\,|\,y)$ the posterior density of β given that they are outliers. The presence of outliers is thus handled automatically. If a subset is particularly discrepant, the corresponding weight $w_{(r)}$ will be large and our ideas about β will allow for the discrepancies.

In each of the following three models, $p_{(r)}(\beta\,|\,y)$ turns out to be a p-variate Student's t distribution with mean $\beta_{(r)}$, dispersion matrix $B^{-1}{}_{(r)}$ and degrees of freedom $\upsilon_{(r)}$, say. It is the different ways in which they treat the suspect observations in arriving at the quantities, especially $\beta_{(r)}$, that is interesting. The posterior weights $w_{(r)}$ are complex, but for a *given* number of outliers we always get $w_{(r)}$ inversely proportional to some power of $s^2_{(r)}$, a kind of "residual sum of squares" from the analysis allowing for outliers.

We shall refer throughout to the standard least-squares values

$$\hat{\beta} = (\chi'\chi)^{-1}\chi' y$$
$$s^2 = (y-\chi\hat{\beta})'(y-\chi\hat{\beta})$$

Box and Tiao (1968) first considered this problem and their model (BT) assumes that each observation has probability 1-α of being generated by the usual linear model and small probability α of coming from the same model but with error variance $k^2\sigma^2$ instead of just σ^2. They took k and α as known and used the usual improper uniform prior on β and log σ.

Here

$$\hat{\beta}_{(r)} = [\chi'_{(n-r)}\chi_{(n-r)} + k^{-2}\chi'_{(r)}\chi_{(r)}]^{-1}[\chi'_{(n-r)}y_{(n-r)} + k^{-2}\chi'_{(r)}y_{(r)}]$$
$$s^2_{(r)} = [y_{(n-r)} - \chi_{(n-r)}\hat{\beta}_{(r)}]'[y_{(n-r)}-\chi_{(n-r)}\hat{\beta}_{(r)}] + k^{-2}[y_{(r)}-\chi_{(r)}\hat{\beta}_{(r)}]'[y_{(r)}-\chi_{(r)}\hat{\beta}_{(r)}]$$
$$\upsilon_{(r)} = n\text{-}p$$
$$B_{(r)} = (n\text{-}p)\,[\chi'_{(n-r)}\chi_{(n-r)} + k^{-2}\chi'_{(r)}\chi_{(r)}]/s^2_{(p)}$$

and $w_{(r)} \propto \{\alpha/k(1\text{-}\alpha)\}^r\,|\chi'_{(n-r)}\chi_{(n-r)} + k^{-2}\chi'_{(r)}\chi_{(r)}|^{-1/2}\,s_{(r)}{}^{-(n-p)}$

Each suspected outlier is thus dealt with by dividing the y value and the corresponding row of the χ matrix through by k and then doing the usual least-squares analysis on this new dataset.

Additive, rather than multiplicative, contamination of the data was considered by Abraham and Box (1978). Their model (AB) was

$$y = \chi\beta + \delta Z + e$$

where Z is a vector each of whose n elements has probability α of being 1 and 1-α of being 0. The amount of contamination δ is thus assumed to be the same for each outlier. Any particular Z vector written $Z_{(r)}$ say, corresponds to a subset of observations being outliers.

Taking α known and improper uniform prior on β,δ and log σ gives

$$\hat{\beta}_{(r)} = [\chi' V_{(r)}\chi]^{-1}\chi' V_{(r)} y \text{ where } V_{(r)} = I\text{-}r^{-1}Z_{(r)}Z'_{(r)}$$
$$s^2_{(r)} = [y\text{-}\chi\hat{\beta}_{(r)}]'\,V'_{(r)}[y\text{-}\chi\hat{\beta}_{(r)}]$$
$$\upsilon_{(r)} = n\text{-}p\text{-}1$$
$$B_{(r)} = \frac{n\text{-}p\text{-}1}{s^2_{(r)}}\,\chi' V_{(r)}\chi$$

and $w_{(r)} \propto [\alpha/(1\text{-}\alpha)]^r\,r^{1/2}\,|\chi' V_{(r)}\chi|^{-1/2}\,s_{(r)}{}^{-(n-p-1)}$

This model thus copes with outliers by doing a weighted least squares analysis using the weighting matrix $V_{(r)}$.

Guttman, Dutter and Freeman (1978) consider additive contamination in a rather different way. Their model (GDF) is

$$y = \chi\beta + a + e$$

where a is a vector exactly r of whose elements are non-zero. They assume the value of r is known, but hedge their bets by doing separate analyses for r = 0,1,2, ... The non-zero elements of a are not forced to be equal, but form r extra unkown parameters which are duly given a uniform improper prior along with β and log σ. We now get

$$\hat{\beta}_{(r)} = [\chi'_{(n-r)}\chi_{(n-r)}]^{-1}\chi'_{(n-r)}y_{(n-r)}$$

$$s^2_{(r)} = [y_{(n-r)}-\chi_{(n-r)}\hat{\beta}_{(r)}]'[y_{(n-r)}-\chi_{(n-r)}\hat{\beta}_{(r)}]$$

$$v_{(r)} = n\text{-}p\text{-}r$$

$$B_{(r)} = \frac{n\text{-}p\text{-}r}{s^2_{(r)}}\chi'_{(n-r)}\chi_{(n-r)}$$

and $w_{(r)} \propto |\chi'_{(n-r)}\chi_{(n-r)}|^{-1/2}s_{(r)}{}^{-(n-p-r)}$

The effect of allowing "totally unknown" amounts of contamination is therefore the dramatic one of dropping suspect observations completely and doing a least-squares analysis on the others.

2. DARWIN'S DATA

All these papers apply their results to the famous set of data due to Darwin quoted by Fisher (1960) and eternally popular with students of outliers.

Here the $n = 15$ observations are

-67 -48 6 8 14 16 23 24 28 29 41 49 56 60 75

and β is the unknown population mean, so $p = 1$ and χ is a column vector of ones.

Box and Tiao display the posterior density of β when $\alpha = .05$ and $k=5$. In identifying outliers, the largest posterior probabilities are as follows:

Outliers	:	None	y_1 and y_2	y_1 only	y_2 only	y_{15} only
Prior prob	:	.463	.0013	.024	.024	.24
Posterior prob	:	.462	.190	.175	.036	.016

If we condition on a fixed number of outliers, we have

$$w_{(r)} \propto S_{(r)}{}^{-(n-1)}$$

where $S^2_{(r)} = \Sigma_{(n-r)}[y_i-\hat{\beta}_{(r)}]^2+k^{-2}\Sigma_{(r)}[y_i-\hat{\beta}_{(r)}]^2$

and $\hat{\beta}_{(r)} = \dfrac{\Sigma_{(n-r)}y_i+k^{-2}\,\Sigma_{(r)}y_i}{n-r\,+\,k^{-2}r}$

in obvious notations.

The largest of these conditional probabilities, for $r=1$ and 2, are given in the columns headed BT of table 1.

TABLE 1

Posterior probabilities, given one or two outliers, for Darwin's data

One outlier			Two outliers					
Observation number	BT	AB = GDF	Observation Pair	BT	Observation Pair	GDF	Observation pair	AB
1	.588	.579	1,2	.785	1,2	.751	1,2	.646
2	.120	.120	1,15	.037	1,15	.037	1,3	.002
15	.053	.054	1,14	.016	1,14	.017	1,4	.002
14	.030	.031	1,13	.013	1,13	.014	1,5	.001
13	.027	.028	1,12	.011	1,12	.012	1,6	.001
12	.023	.023	1,3	.010	1,3	.011	14,15	.001
11	.020	.020	1,4	.010	1,4	.011	13,15	.001
3	.018	.019	1,11	.009	1,11	.010	1,7	.001
4	.018	.019	1,6	.008	1,6	.010		

A sensitivity analysis showed that the posterior mean and variance of β are hardly affected by large changes in the value of k. While changes in α are rather more crucial, there is still a fair amount of robustness and the results do not vary much as α ranges between .03 and .07.

In the Abraham and Box model

$$\chi' V_{(r)}\chi = n\text{-}r, \qquad \hat{\beta}_{(r)} = \bar{y}_{(n-r)}$$

$$s^2_{(r)} = \Sigma_{(r)}[y_i-\bar{y}_{(r)}]^2+\Sigma_{(n-r)}[y_i-\bar{y}_{(n-r)}]^2$$

The first term here clearly arises as a consequence of the assumption of the same δ for each outlier, $\bar{y}_{(r)}$ being the natural estimate of $\beta+\delta$.

Conditionally on r,

$$w_{(r)} \propto S_{(r)}{}^{-(n-2)}$$

Note that, since $p=1$, all suspect observations are ignored in forming $\hat{\beta}_{(r)}$ but contribute towards $S^2_{(r)}$ except when $r=1$. In that case these results coincide with those of the GDF model.

Abraham and Box give the posterior density of β for a range of α values, do a sensitivity analysis on the mean and variance of β as α changes, and quote conditional posterior probabilities $w_{(r)}$ for $r=1$ and 2, reproduced here in table 1.

In the Guttman, Dutter and Freeman model,

$$\hat{\beta}_{(r)} = \bar{y}_{(n-r)}$$
$$S^2_{(r)} = \Sigma_{(n-r)}[y_r\text{-}\bar{y}_{(n-r)}]^2$$
and $w_{(r)} \propto S_{(r)}^{-(n-1-r)}$

these latter being inherently conditional on fixed r.

As Table 1 shows for only one outlier all three models agree on observation 1(-67) as being by far the most likely candidate. All the central observations from 6 to 29 get almost identical posterior probabilities, as dropping any one of them makes very little difference to the sum of squares about the mean. For two outliers, however, the Abraham-Box model diverges from the others in that it cannot encompass the possibility that outliers might occur in both tails of the distribution. It also gives less posterior weight to the most obvious pair (-67, -48) and spreads the posterior probability pretty uniformly over all except three pairs. The model is clearly not a good one for identifying outliers and so must necessarily be weak at providing robust estimates of β under some circumstances.

3. A 2^5 FACTORIAL EXPERIMENT

John (1978) discussed the results of a 2^5 factorial experiment in two blocks with the ABCDE interaction confounded. Visual inspection of a plot of residuals against fitted values suggests that there might be two outliers. Having derived a suitable test statistic and simulated its sampling distribution, a significance level α = .117 was obtained, from which it was concluded that there were not two outliers. Had a test for only one outlier been performed, however, the result would have been significant with α = .044.

Besag (1979) reports that a robustified regression analysis, using Tukey's "exploratory data" approach, clearly shows the presence of one outlier, not two.

An analysis using the GDF model fitting main effects and first-order interactions confirms this approach. Table 2 shows that assuming one outlier gives posterior probability .734 to one of the observations, whereas the most likely pair only gets probability .147. The posterior mean of β changes markedly as we change from 0 to 1 outlier but hardly at all when we progress to 2 outliers. The sum of the posterior variances of the elements of β is again least for one outlier.

TABLE 2

Data on 2^5 factorial experiment, from John (1978)

DATA

(1)	1.4	d	5.0	e	1.7	de	9.5
a	1.2	ad	9.0	ae	2.0	ade	5.9
b	3.6	bd	12.0	be	3.1	bde	12.6
ab	1.2	abd	5.4	abe	1.2	abde	6.3
c	1.5	cd	4.2	ce	1.9	cde	8.0
ac	1.4	acd	4.4	ace	1.2	acde	4.2
bc	1.5	bcd	9.3	bce	1.0	bcde	7.7
abc	1.6	abcd	2.8	abce	1.8	abcde	6.0

POSTERIOR PROBABILITIES

One outlier		Two outliers	
.734	ad	.147	ad, acd
.098	d	.090	d, ad
.010	bcd	.050	ad, abcd
.009	bce	.047	ad, bcde
.008	abcd	.040	ad, ace
.008	abcde	.040	ad, abce

POSTERIOR MEAN AND VARIANCE OF β

	MEAN			VARIANCE		
N° Outliers						
	0	1	2	0	1	2
(1)	4.36	4.24	4.22	.087	.074	.082
a	-0.89	-1.04	-1.09	.087	.065	.067
b	0.46	0:58	0.60	.087	.074	.084
c	-0.71	-0.58	-0.58	.087	.074	.073
d	2.66	2.53	2.51	.087	.074	.082
e	0.27	0.40	0.42	.087	.074	.084
ab	-0.64	-0.49	-0.44	.087	.065	.067
ac	0.16	0.31	0.32	.087	.065	.063
ad	-0.63	-0.79	-0.83	.087	.066	.068
ae	-0.17	-0.01	0.03	.087	.066	.068
bc	-0.15	-0.28	-0.27	.087	.074	.071
bd	0.29	0.41	0.44	.087	.074	.085
be	-0.12	-0.25	-0.27	.087	.074	.085
cd	-0.49	-0.36	-0.36	.087	.074	.072
ce	0.05	-0.08	-0.07	.087	.074	.072
de	0.24	0.36	0.38	.087	.074	.084
			Total	1.393	1.139	1.206

4. HOW MANY OUTLIERS?

While this question is less interesting that the main one of the unkown value of β, there are some examples in which it is important to have a fairly clear answer. A central laboratory receiving routine radioimmunoassay readings from a number of medical centres, for example, needs not only to allow for outliers during analysis of the collected data, but also to note which centres are consistently producing relatively large numbers of outliers so that their experimental techniques can be kept up to scratch.

In answering the question we always have to be careful not to compare models with different numbers of parameters since if we do, using improper priors of different dimensionality, the posterior probabilities we obtain will be meaningless. Box and Tiao can safely derive the probabilities we quote in section 2 of 0, 1 or 2 outliers since their model always has $p+1$ parameters, independent of r.

To attempt to do the same for the AB model would be disastrous, however, as this has $p+1$ parameters when $r \neq 0$ but only p when $r=0$. The GDF model carries this problem further as each new outlier adds a new unknown parameter. A naive attempt to apply the formal analysis would merely lead to nearly all the posterior probability being heaped onto the largest number of outliers considered, since it can never do any harm to add more parameters. There is much current discussion about what is a fair penalty to expect a complex model to pay when comparing it with a parsimonious one, but as yet no general agreement. Akaike's (1973) very popular AIC criterion cannot be used here as the likelihood functions of all these models are themselves sums of 2^n or, (for GDF) ${}^nC_r$ terms each of which are products of normal distributions, so that the maximum likelihood estimates needed to evaluate the maximum of the likelihood functions are impossible to find analytically.

We propose here to sidestep this general question by pursuing the GDF model using proper priors throughout. While this automatically removes all doubt about whether the answers are right, it simultaneously introduces the need for a sensitivity analysis to see to what extent those answers depend on the particular priors used.

We first assign prior probability π_r to there being r outliers ($r=0, 1, \ldots, n$, $\Sigma\pi_r = 1$) and refer to this as "model r". Within this model we look at all ${}^nC_r$ possible partitions of the observations and assign prior probability $\pi_{(r)}$ to a particular subset being the outliers ($\Sigma_{(r)}\pi_{(r)} = 1$). Conditionally on this we now assign prior densities for the unknown parameters. We take β given σ^2 as p-variate normal with mean b_o and dispersion matrix $\sigma^2 B_o$ and a given σ^2 as r-variate normal with mean a_o and dispersion matrix $\sigma^2 A_o$. Finally we take $\upsilon v/\sigma^2$ as chi-square on υ degrees of freedom. We suppress the subscript (r) on the quantities b_o, B_o, a_o, A_o, υ and v partly for simplicity but mainly because in practice it is difficult to envisage how these could depend on the particular subset being considered.

Conditional on any given subset being outliers, the posterior for β is Student's t with

$$\hat{\beta}_{(r)} = B_{(r)}{}^{-1} d_{(r)}$$

$$\nu_{(r)} = n + \nu$$

$$B_{(r)} = B_0{}^{-1} + \chi'_{(n-r)}\chi_{(n-r)} + \chi'_{(r)}(A_0 + \mathrm{I})^{-1}\chi_{(r)}$$

and $$w_{(r)} \propto |B_{(r)}|^{-1/2} D_{(r)}{}^{-(n+\nu)/2} |A_0+\mathrm{I}|^{-1/2} \pi_{(r)}$$

where $$d_{(r)} = B_0^{-1}b_0 + \chi'_{(n-r)}y_{(n-r)} + \chi'_{(r)}(A_0+\mathrm{I})^{-1}(y_{(r)}-a_0)$$

and $$D_{(r)} = b_0' B_0^{-1} b_0 + \nu v + y'_{(n-r)} y_{(n-r)} + (y_{(r)}-a_0)'(A_0+\mathrm{I})^{-1} (y_{(r)}-a_0) - d_{(r)}' B_{(r)}^{-1} d_{(r)}.$$

The posterior mean of β, for example, given model r, is

$$E_r(\beta|y) = \frac{\Sigma_{(r)} w_{(r)} \hat{\beta}_{(r)}}{\Sigma_{(r)} w_{(r)}}$$

the sums being over all ${}^nC_r$ possible partitions into r and $n-r$ observations.

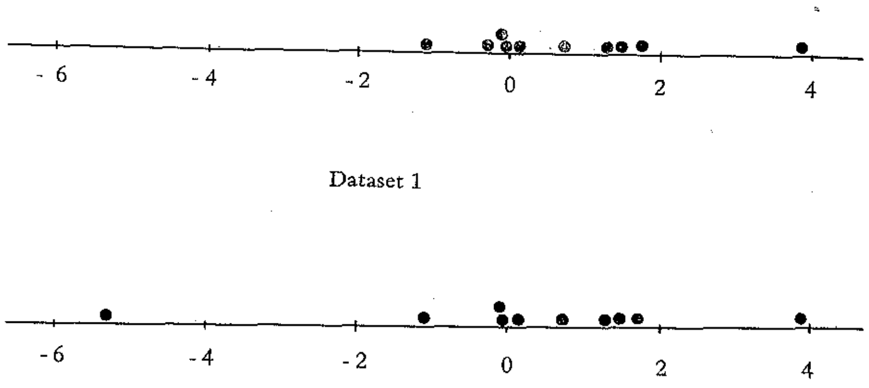
The prior probability π_r that model r is true is changed into the posterior probability

$$\pi_r' \propto \pi_r \Sigma_{(r)} |B_{(r)}|^{-1/2} D_{(r)}{}^{-(n+\nu)/2} |A_0+\mathrm{I}|^{-1/2} \pi_{(r)} .$$

We note that the effect of taking very vague, but proper, priors on a within each model is to throw all the posterior weight on $r = 0$, the simplest model, since then $|A_0+\mathrm{I}|^{-1/2}$ decreases geometrically in r. Like many "modern" results this simply rediscovers the work of Jeffreys (1961). The contrast with improper priors which put most posterior weight on the most complicated model is, however, so stark as to be worth mentioning again.

5. SOME TEST DATA

The trouble with proper priors, of course, is actually specifying them. A sensitivity analysis is essential to establish the influence of fairly large changes in the priors on the posterior statements. Darwin's data are not a very suitable set for seeing how well the above results perform as it is not at all clear how many outliers there really are. Accordingly, 10 random observations from N (0,1) were taken. Dataset 1 was formed by adding 4 to one of the observations, and dataset 2 by further adding -5 to another, see fig. 1. Any self-respecting method ought to be able to get the right answers in such clear-cut cases.



Dataset 1

Dataset 2

FIGURA 1

We assigned equal probability $\frac{1}{4}$ to the number of outliers r being 0,1,2 or 3 and assumed that within model r all ^nC_r subsets of r outliers were equally likely. We also took the elements of a to be identically and independently distributed $N(a_0, A_0\sigma^2)$, where a_0 and A_0 are now scalars, the same for each value for r.

Thinking firstly of dataset 1, we might agree that the "right" priors are

$$\beta \sim N(0, \sigma^2)\,,\; a \sim N(4, A_0\sigma^2)\,,\; 8\sigma^{-2} \sim \chi^2_{10}\,.$$

The last of these gives prior mode $= \frac{2}{3}$, mean $= 1$ and variance $\frac{1}{3}$ for σ^2. We allow A_0 to vary between 10^{-4} and 10^4 since we know that this will crucially affect the answers. These come out to be as in fig. 2(a), that is that we get the clear, correct message that there is one outlier so long as A_0 is not too large. When we use the same priors on dataset 2, however, we get fig 2(b), which completely fails to detect the two outliers. This is hardly surprising since the prior on a is now highly inappropriate.

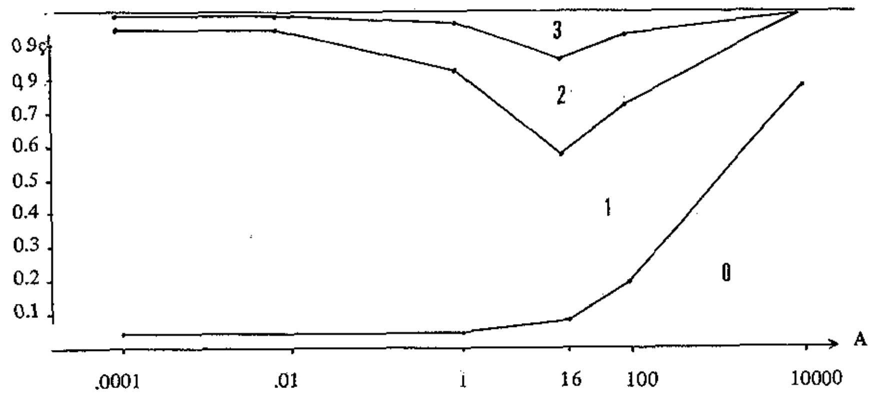


FIGURA 2a

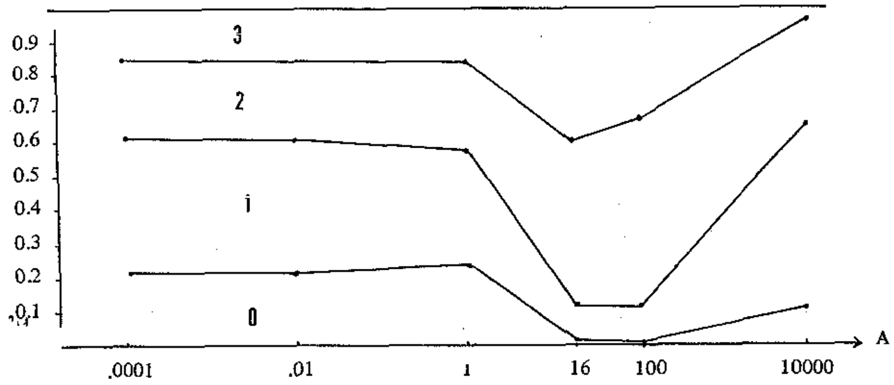


FIGURA 2b

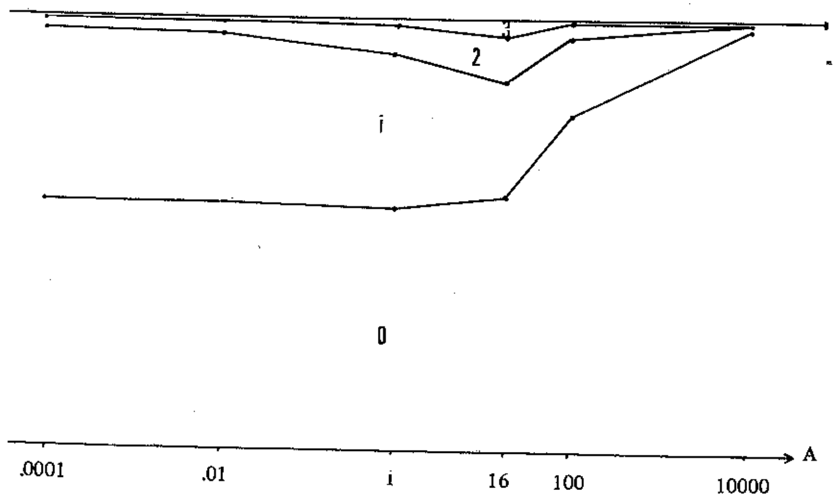


FIGURA 3a

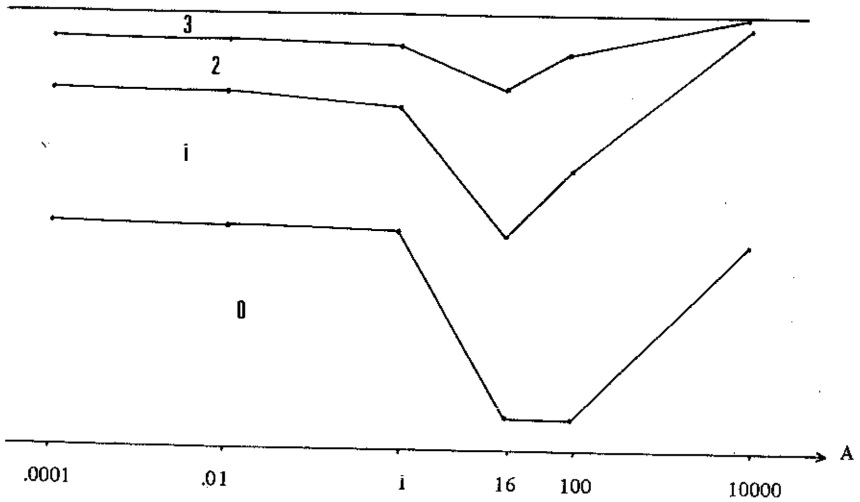


FIGURA 3b

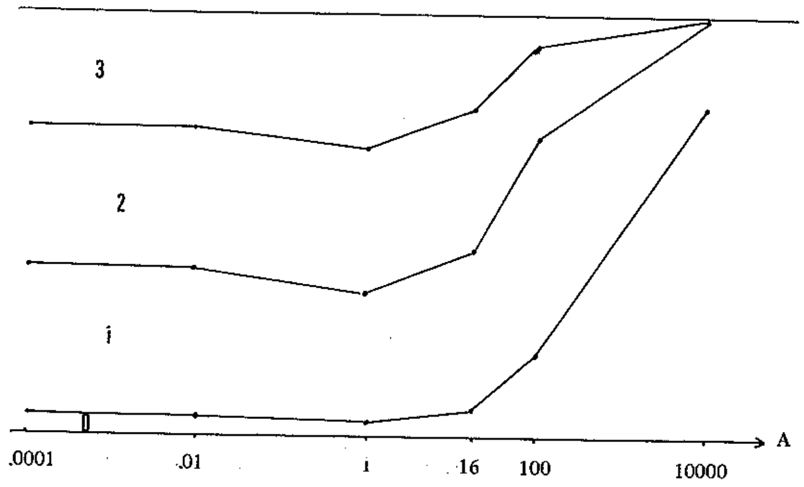


FIGURA 4a

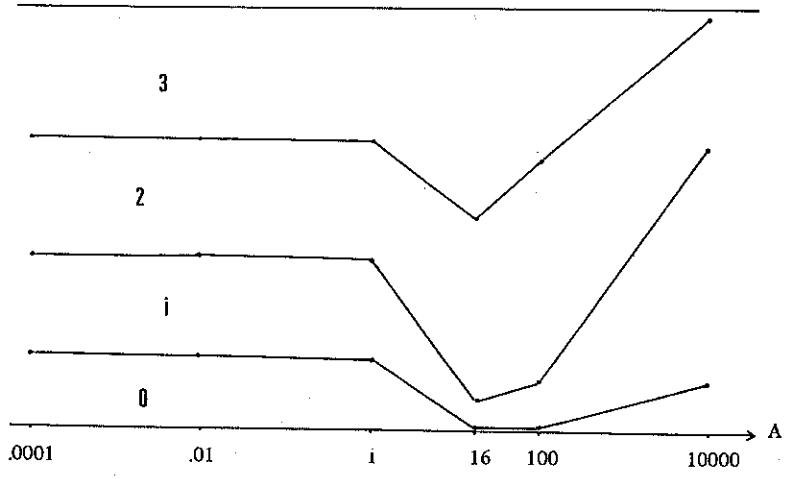


FIGURA 4b

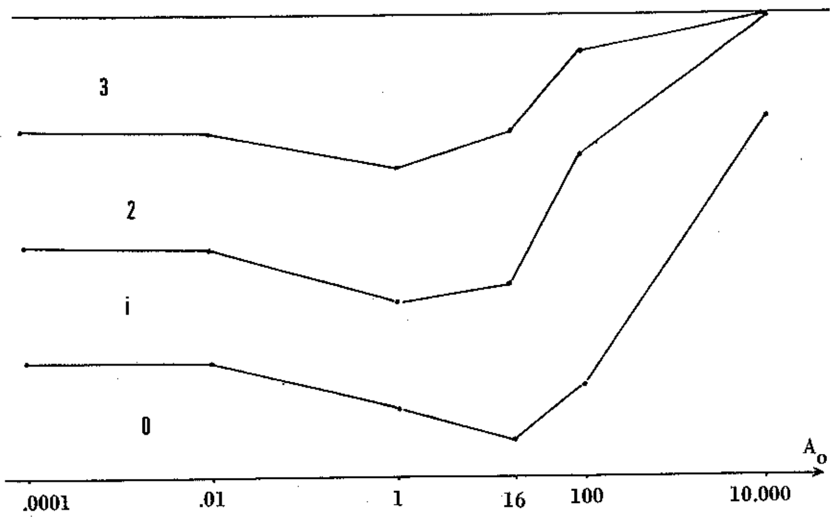


FIGURA 5a

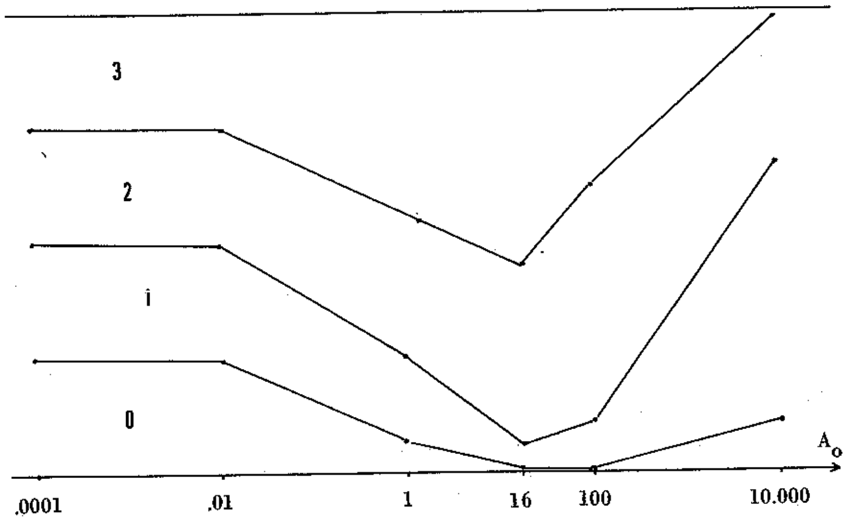


FIGURA 5b

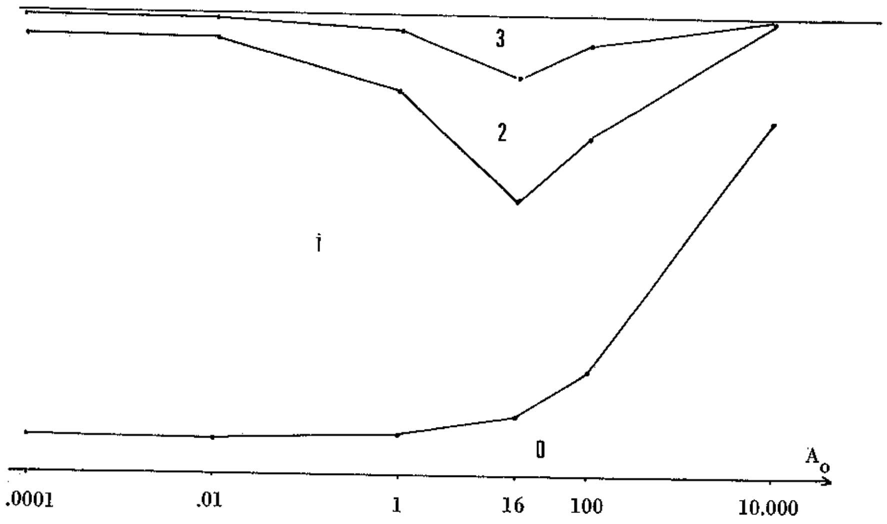


FIGURA 6a

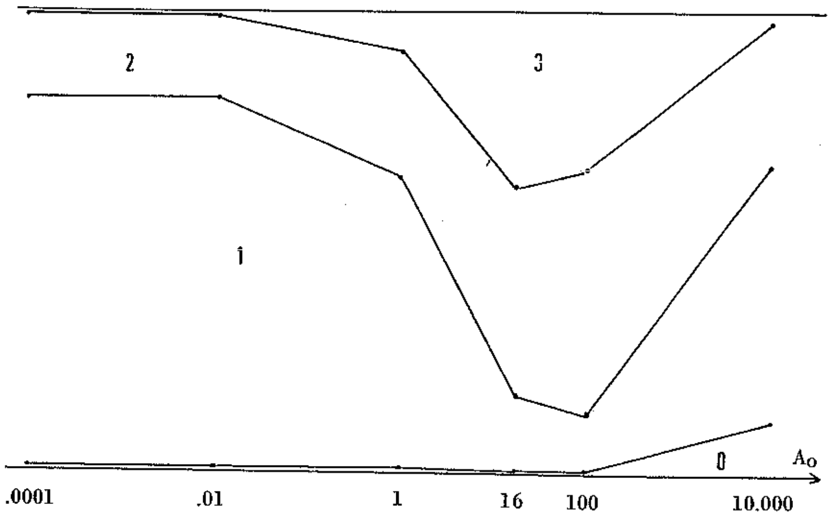


FIGURA 6b

Returning to dataset 1, if we use $N(0,100\sigma^2)$ for β we get virtually the same result as we do also using $N(5,100\sigma^2)$, but $N(5,\sigma^2)$ gives fig 3(a) in which we end up quite sure that $r = 0$ or 1 but fail to distinguish clearly between them. Replacing the prior for σ^2 by $2\sigma^{-2} \sim \chi^2_1$, having the same mode of $\frac{2}{3}$ but infinite mean and variance, makes the results slightly less sharp but substantially unaltered. If we get the prior mean of a wrong, though, the results are disastrous. Figs 4(a) and 5(a) show the effects of taking $a_0 = 2$ and $a_0 = 0$ respectively. The latter can be thought of as the closest the GDF model can get to the Box-Tiao philosophy. Not surprisingly, small values of A_0 give posterior probabilities exactly the same as the prior ones. As A_0 increases the probability of one outlier starts to build up but doesn't get near to being decisive before the inevitable slide towards no outlier sets in.

Turning to dataset 2, the corresponding results in figs 2(b), 3(b), 4(b) and 5(b) are all disappointing, especially the last. Taking zero prior mean with a large prior variance for a might have promised to model successfully the occurrence of "two-sided" outliers, but that large prior variance proves its downfall. A preference for two outliers is just starting to show when increasing A_0 pushes the probabilities down towards one and zero outliers. Another hopeful prior might be the mixture $\frac{1}{2}N(4, A_0\sigma^2) + \frac{1}{2}N(-4, A_0\sigma^2)$ but fig 6 shows that while this continues to pick out one outlier successfully, it has no better luck with two than any of its predecessors.

Perhaps this poor performance is not so disgraceful as it seems at first blush. Gentle (1979) reported simulation studies of his proposed frequentist-based outlier detection procedures. For twenty observations with p (the dimension of β) $= 2$ two outliers were correctly identified only 28% of the time. This rose to 74% for 40 observations and 82% with 60. One hope for our approach, then, might be to increase n in this fashion, but this would immediately create the usual combinatorial explosion and become prohibitively expensive on computer time. By their very nature all three models can only be used with small sample sizes unless a maximum of two outliers is contemplated.

6. DISCUSSION

The GDF model using proper priors can tentatively be claimed to be insensitive to choice of prior on σ^2 and β, so long as a too-precise wrong value of b_0 is not used. It is, however, very sensitive to choice of a_0 and care must be taken not to set A_0 too large. There is also at present no known prior structure that permits large positive and negative contaminations to show themselves simultaneously. On the other hand there is no set of improper priors that would generally be agreed to be appropriate for this problem. Perhaps some of the other papers at this conference will propose a way forward but it might be that attempts like the AIC criterion to produce a standard way of answering a wide variety of questions regardless of their different contexts are doomed to failure.

Although the question 'How many outliers' may easily be dismissed as an unimportant one, so long as robust inferences about β and σ^2 are possible, I prefer to see it as just one manifestation of the model discrimination problem that is the biggest current challenge to Bayesian statisticians.

REFERENCES

ABRAHAM, B. and BOX, G.E.P. (1978). Linear models and spurious observations. *Appl. Statist.* **27**, 131-8.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (B.N. Petrov and F. Csaki, eds.) 267-281, Budapest, Akademia Kiado.

BESAG, J. (1979). Exploratory data analysis. *Invited paper to R.S.S. Conference*, Oxford, April 2-6.

BOX, G.E.P. and TIAO, G.C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55**, 119-29.

FISHER, R.A. (1960). *The design of experiments* (7th ed.) Oliver and Boyd: Edinburgh.

GENTLE J.E. (1978). Testing for outliers in linear regression. In *Contributions to survey sampling and applied statistics* (H.A. David ed.) 223-233. New York: Academic Press.

GUTTMAN, I., DUTTER, R. and FREEMAN, P.R. (1978). Care and handling of univariate outliers in the general linear model to detect spuriosity - a Bayesian approach. *Technometrics* **20**, 187-193

JEFFREYS, H. (1961). *Theory of probability*. Oxford:University Press.

JOHN, J.A. (1978). Outliers in factorial experiments. *Appl. Statist.* **27**, 111-9.

Sampling Inference, Bayes' Inference, and Robustness in the advancement of learning*

G.E.P. BOX

University of Wisconsin-Madison

SUMMARY

Scientific learning is seen as an iterative process employing Criticism and Estimation. Sampling theory use of predictive distributions for model criticism is examined and also the implications for significance tests and the theory of precise measurement. Normal theory examples and ridge estimates are considered. Predictive checking functions for transformation, serial correlation, and bad values are reviewed as is their relation with Bayesian options. Robustness is seen from a Bayesian view point and examples are given. The bad value problem is also considered and comparison with M estimators is made.

Keywords: ITERATIVE LEARNING; MODEL BUILDING; INFERENCE; BAYES THEOREM; SAMPLING THEORY; PREDICTIVE DISTRIBUTION; DIAGNOSTIC CHECKS; TRANSFORMATIONS; SERIAL CORRELATION; BAD VALUES; OUTLIERS; ROBUST ESTIMATION.

Scientific method is a process of guided learning in which accelerated acquisition of knowledge relevant to some question under investigation is achieved by a hierarchy of iterations in which induction and deduction are used in alternation.

This process employs a developing model (or series of models implicit or explicit) against which data can be viewed. At any given stage of the investigation, the current model approximates relevant aspects of the studied system and motivates the acquisition of further data as well as its analysis. By the use of a prior distribution it is possible to represent some aspects of such a model as if they were completely known and others as if they were more or less unknown.

Now parsimony requires that, at any given stage, the model is no more complex than is necessary to achieve a desirable degree of approximation and since each investigation is unique we cannot be sure in advance that any model we postulate will meet this goal. Therefore, at the various points in our investigation where data analysis is required, two types of inference are involved: *model criticism* and *parameter estimation*. To effect the latter, conditional on the plausibility of the model, and given the data, we can, using Bayes' Theorem, deduce posterior distributions for unknown parameters and so make inferences about them. But, before we can rely on such conditional deduction, we ought logically to check whether the model postulated accords with the data at all and, if not, consider how it should be modified. In practice, this question is usually investigated by inspecting residuals, by other informal techniques, and sometimes by making formal tests of goodness of fit. In any case this inferential procedure of model criticism whereby the need for model modification is induced, is ultimately dependent on sampling theory argument. These principles may be formalized by an appropriate analysis of Bayes' formula.

TWO COMPLEMENTARY FACTORS FROM BAYES FORMULA

If we accept the prior probability distribution of parameters θ as an essential part of a model then all aspects of the model, hypothesized at some particular stage of an investigation, are contained in the joint density obtained by combining the likelihood and the prior

$$p(\mathbf{y},\theta\,|\,M) = p(\mathbf{y}\,|\,\theta,M)\,.\,p(\theta\,|\,M) \qquad (1)$$

where $|M$ is understood to indicate conditionality on some aspect of the model and $\mathbf{y}$ is the data vector.

This joint distribution which is a comprehensive statement of the model can also be factored as

* This is an extended abstract of the paper "Sampling and Bayes' Inference in Scientific Modelling and Robustness" later read before the Royal Statistical Society at Cardiff, Wales, on May 15th, 1980 and published in *J. Roy. Statist. Soc. A*, **143**.

$$p(\mathbf{y},\theta\,|\,M) = p(\theta\,|\,\mathbf{y},M)p(\mathbf{y}\,|\,M) \tag{2}$$

and can be computed before any data becomes available. In particular the second factor on the right

$$p(\mathbf{y}\,|\,M) = \int p(\mathbf{y}\,|\,\theta,M)p(\theta\,|\,M)d\theta \tag{3}$$

which is the *predictive* distribution, may be so calculated. It is the distribution of the totality of all possible samples that could occur if the model M were true.

When an actual data vector $\mathbf{y}_d$ becomes available

$$p(\mathbf{y}_d,\theta\,|\,M) = p(\theta\,|\,\mathbf{y}_d,M)p(\mathbf{y}_d\,|\,M) \tag{4}$$

The first factor on the right is then Bayes' posterior distribution of θ given $\mathbf{y}_d$

$$p(\theta\,|\,\mathbf{y}_d,M) = k_d p(\mathbf{y}_d\,|\,\theta,M)p(\theta\,|\,M) \tag{5}$$

and the second factor

$$p(\mathbf{y}_d\,|\,M) = \int p(\mathbf{y}_d\,|\,\theta,M)p(\theta\,|\,M)d\theta = k_d^{-1} \tag{6}$$

is the predictive density associated with the data set $\mathbf{y}_d$ actually obtained.

If the model is to be believed, then the posterior distribution $p(\theta\,|\,\mathbf{y}_d,M)$ allows all relevant estimation inferences to be made about θ. However even if the model were totally incorrect, this could not be shown by any abnormality in this factor which is conditional on *both* data and model specification. However, plausibility or otherwise of obtaining such a sample *if the model were appropriate* may be assessed by reference of the density $p(\mathbf{y}_d\,|\,M)$ to the predictive reference distribution $p(\mathbf{y}\,|\,M)$. An unusually small value of $p(\mathbf{y}_d\,|\,M)$ as measured by $Pr\{p(\mathbf{y}\,|\,M) < p(\mathbf{y}_d\,|\,M\}$ casts doubt on the appropriateness of the model M.

Now $p(\mathbf{y}\,|\,M)$ is an n-dimensional distribution and it will usually be true that if the model is inadequate it is most likely to be deficient in certain directions associated with unusual values of certain specific functions $g_i(\mathbf{y})$ of the data. Examples of such functions are sample averages, variances, moment coefficients, coefficients of serial correlation, and measures of standardized deviations from a norm. In every case the appropriate reference distribution to which the realized statistic $g_i(\mathbf{y}_d)$ should be referred is the distribution $p\{g_i(\mathbf{y}\,|\,M)\}$, when the model M is true, derived by appropriate integration of

$p(\mathbf{y}\,|\,M)$.

In practice, criticism or diagnostic checking of the model is often conducted by visual inspection of residual displays and other more sophisticated plots. But such a process, although it is informal, still, it seems to me, falls within the logical framework described above. The statistician is looking for "features" in the data which would be surprising or "unusual" if the model M were true. Such a feature can be described by a function $g(\mathbf{y}_d)$ and its unusualness, if formalized, would have to be measured by reference to $p\{g(\mathbf{y}\,|\,M)\}$.

DIAGNOSTIC CHECKS AND ROBUSTIFICATION

A question which confronts the statistician at every stage of an investigation is "How complex a model should I use?" An apparently different question is "Should I use a robust procedure?", but I will argue that this is subsumed by the broader question. The possibilities for model elaboration are of course limitless. For instance a commonly used model assumes errors to be Independently, Identically and Normally distributed (IIN). It is easy to imagine a sequence of fallback models which begin like this

$$\begin{array}{ccccccccc} M_0 & \rightarrow & M_1 & \rightarrow & M_2 & \rightarrow & M_3 & \rightarrow & \ldots. \\ \mathrm{IIN} & & \mathrm{IIN} & & \not{\mathrm{I}}\mathrm{IN} & & \not{\mathrm{I}}\not{\mathrm{I}}\mathrm{N} & & \end{array}$$

Obviously compromise is necessary; for on the one hand, simpler models can allow better scientific understanding and better estimation, while on the other hand, more complex ones can be, but need not be, closer to the truth.

This raises the problem of where should the compromise be made. Suppose some deviation from an "ideal" model M_0 can be parameterized by a *discrepancy parameter* β or a vector of such parameters. In each case there are two ways to handle the possible model discrepancy, depending on whether the parameter β is omitted from, or included in, the model. We call these *diagnostic checking and robustification.*

Diagnostic checking. If the discrepancy parameter is omitted from the model then an appropriate diagnostic check can be made. Formally this may be done by referring some suitable function $g_\beta(\mathbf{y})$ of the data to a reference distribution derived from the predictive distribution $p(y\,|\,M_0)$.

Robustification. If the discrepancy parameter is included then robust estimation of θ is provided by the posterior distribution

$$p(\theta\,|\,y) = \int p(\theta\,|\,\beta,\mathbf{y})p(\beta\,|\,\mathbf{y})d\beta \tag{7}$$

If we write

$$p_u(\beta\,|\,\mathrm{y}) = p(\beta\,|\,\mathrm{y})/p(\beta) \qquad (8)$$

$$p(\theta\,|\,\mathrm{y}) = \int p(\theta\,|\,\beta,\mathrm{y})p_u(\beta\,|\,\mathrm{y})p(\beta)d\beta \qquad (9)$$

In the last expression

(i) $p(\beta)$ can be chosen to represent approximately the probability of occurrence of different values of β feared in the real world

(ii) the function $p_u(\beta\,|\,y)$ is a pseudo-likelihood which reflects information about β supplied by the data

(iii) considered as a function of $\beta, p(\theta\,|\,\beta,\mathrm{y})$ reflects the sensitivity of estimation to the choice of the discrepancy parameter.

Numerous authors (Huber, Tukey, Andrews, Hampel, etc.) have proposed various methods of robust estimation relying on the empirical modification of classical estimation procedures. It seems more logical to me to modify the model which is presumably at fault rather than the method of estimation which is not. Furthermore this has the advantage of clearly revealing the assumptions which are being made.

The primary candidates for inclusion in the model (robustification) *reflect features which might easily elude diagnostic checks and could then invalidate subsequent analysis.* But, however the model may be elaborated it will still be necessary to apply diagnostic checks (for example, to study residuals). Thus model criticism using sampling theory and parameter estimation using Bayes theorem fill different but necessary roles in the scientific iteration.

DISCUSSION

W.F. EDDY (*Carnegie-Mellon University*):

I predict that by the end of this century the religious cult of Pure Bayesian Statistics (PBS) will die. There will be no martyrs. Righteousness is not the question; God will not decide in favor of incoherence and destroy Las Fuentes as he destroyed Sodom and Gomorrah. PBS will die the death of the buggy whip, through disuse.

Lest I be misunderstood, by PBS, I mean the belief that finding the distribution of unknown parameters *conditional on the data assuming the truth of the model is the objective of statistics.* The fundamental difficulty with PBS is that all inference is based on the truth of the model. And despite disclaimers I doubt that any practicing statistician believes in the truth of his model.

Professor Box apparently agrees. As I understand his thesis, one first uses sampling theory to find a "true" model and then uses Bayes theory to estimate the parameters in this model. The thrust of his argument is that allowance must be made for the possibility that the model was not sufficiently broad and thus the prior distribution didn't really account for all uncertainty. On the face of it, this is a valuable thought.

However, Professor Box suggests that one should consequently do diagnostic checking. That is, after finding some unusual aspect of the data one should compute a discrepancy function and compare the observed value with the appropriate reference distribution.

This, I believe, is a mistake. Because the particular discrepancy function was chosen after looking at the data the reference distribution will usually suggest the observed value is unusual; but this is exactly the reason we computed the discrepancy function in the first place. Comparing an observed discrepancy to a reference distribution can only be useful for specific *a priori* departures from the model.

This is not to say that examining residuals and computing discrepancies is worthless. On the contrary, there is no substitute for careful residual analysis. Professor Box and I agree on this point and its implication: Model Building/Data Analysis is subjective. Different people see different things in their data and consequently add different parameters to their models.

I don't believe, however, that Professor Box has solved the fundamental dilemma of statistics: How to generalize from the specific data at hand?

Professor Freeman has presented us with a very practical comparison of several "outlier" linear models. I have been intrigued by the models and their implications but I am puzzled about their Bayesian-ness and thus the quotes around "outlier". In common usage, an outlier is an observation which *appears* to be different than the rest of the data (I emphasize *appears* because it is obviously a subjective matter which aspects of the data one examines). Now the Bayesian is compelled to choose his model(s) before seeing the data and thus, it seems to me, is in a quandry as to how to include the outliers in his model since he doesn't yet know which aspects of the data appear to be different. Since the models here are obviously geared to location and scale shifts (slippage outliers) perhaps the outlier-ness of the models is not to be questioned. The solution to my puzzlement may be that Professor Freeman uses the term "outlier" as shorthand for "what a non-Bayesian would call an 'outlier'." Enough philosophy.

By partitioning the data into outliers and non-outliers he writes the posterior distribution of β as

$$p(\beta\,|\,y) = \Sigma w_{(r)}p_{(r)}(\beta\,|\,y)$$

This device has two advantages: first, it allows analysis to proceed conditionally on particular observations being outliers and thus greatly simplifies calculations; second, it allows subsequent inference about which observations are outliers. Professor Freeman considers three specific outlier models: BT, AB, GDF. All three models suppose that the outliers are uniformly distributed over the observations; a more realistic model might distribute them conditionally on X.

The BT model says outliers have the same mean but are scaled by a factor of k. The posterior probabilities ($W_{(r)}$) will be largest when the outliers are observations at one or both extremes. The AB model says all outliers have a different (common) mean.

Consequently, the posterior probabilities will be largest when the outliers are a group of observations at one extreme. The GDF model says outliers each have a different mean. Thus, they are eliminated from the analysis since they contain no information about either β or σ^2; furthermore, the $w_{(r)}$ will be largest when the outliers are two groups, one at each extreme.

All three of the models can be viewed, conditionally on particular observations being outliers, as weighted least squares with the weights depending on the particular outlier model. That is, for all three models

$$\begin{aligned}\hat{\beta}_{(r)} &= (X'V_{(r)}X)^{-1}X'V_{(r)}y\\ s^2_{(r)} &= (y-X\hat{\beta}_{(r)})'V_{(r)}(y-X\hat{\beta}_{(r)})\\ B_{(r)} &= (v_{(r)}/s^2_{(r)})X'V_{(r)}X\end{aligned}$$

and

$$w_{(r)} \propto c_{(r)}\,|X'V_{(r)}X|^{-1/2}s_{(r)}{}^{-v_{(r)}}$$

For simplicity suppose the observations are permuted so the r outliers occur first. Then for the BT model

$$V_{(r)} = \begin{bmatrix} k^{-2}I_r & 0\\ 0 & I_{n-r}\end{bmatrix}$$

$$\nu_{(r)} = n\text{-}p,$$

$$c_{(r)} = \left[\frac{\alpha}{k(1\text{-}\alpha)}\right]^r \text{ and}$$

for the AB model

$$V_{(r)} = \begin{bmatrix} J_r/r & 0\\ 0 & \mathrm{I}_{n-r}\end{bmatrix}$$

where J_r is an r x r matrix of ones.

$$v_{(r)} = n\text{-}p\text{-}1,$$

$$c_{(r)} = (\alpha/(1\text{-}\alpha))^r\,r^{-1/2} \text{ and}$$

for the GDF model

$$V_{(r)} = \begin{bmatrix} 0 & 0\\ 0 & I_{n-r}\end{bmatrix}$$

$$v_{(r)} = n\text{-}p\text{-}r$$

$$c_{(r)} = 1$$

The great advantage is that we can now examine the $V_{(r)}$ to see if we really want to use a particular model; we can quickly examine new proposed outlier models.

I personally find the GDF model somewhat disquieting; completely ignoring extreme observations seems dangerous. An alternative I would prefer is a mixed BT-AB model as follows: With probability α_j each observation has mean $X\beta+\delta_j$ and variance $k^2_j\sigma^2$ for $j=1,2$ and with probability $1\text{-}\alpha_1\text{-}\alpha_2$ each observation has mean $X\beta$. Take $\alpha_1,\alpha_2,k_1,k_2$ known and uniform (improper) priors on β,δ_1,δ_2, and log σ. For r_1 and r_2 outliers, respectively, this yields (in obvious notation)

$$V_{(r_1,r_2)} = \begin{bmatrix} k_1^{-2}(I_{r_1}\text{-}J_{r_1}/r_1) & 0 & 0\\ 0 & k_2^{-2}(I_{r_2}\text{-}J_{r_2}/r_2) & 0\\ 0 & 0 & I_{n-r_1-r_2}\end{bmatrix}$$

$$v_{(r_1r_2)} = n\text{-}p\text{-}2$$

$$c_{(r_1r_2)} = [\alpha_1/k_1(1\text{-}\alpha_1)]^{r_1}\,[\alpha_2/k_2(1\text{-}\alpha_2)]^{r_2}\quad(r_1r_2)^{-1/2}$$

This model uses either location or scale (or both) information from the outliers; only when the r's are one does it reduce to the GDF expedient of ignoring data.

A. O'HAGAN (*University of Warwick*):

Professor Box argues that sampling theory methods are appropriate in diagnostic checking, and I strongly disagree. But whilst elaborating on this, let me say what a pleasure it is to find that he is actually tackling the right problem in basically the right way. The crucial point is the recognition that every statistical analysis, Bayesian or otherwise, is conditional on the truth of its assumptions. Any analysis which goes no further, which does not challenge these assumptions, is incomplete. So Professor Box is right in pointing to a need for procedures for diagnostic checking. And with the accuracy of an experienced data analyst he chooses the right tool, the predictive density $p(\mathbf{y}/M)$. Then inconceivably he uses the tool in entirely the wrong way. There is a perfectly natural Bayesian approach which uses the predictive density but never lapses into the discredited sampling-theory use of tail area probabilities.

Consider the basic model M and an alternative M_1. Conditional on M we obtain the basic posterior density $p(\theta/\mathbf{y}_d,M)$. Or conditional on M_1 we could obtain a different posterior density $p(\theta/\mathbf{y}_d,M_1)$. We now widen the analysis by conditioning on the truth of *either M or M*$_1$. We need extra prior probabilities $P(M/M$ or $M_1)$ and $P(M_1/M$ or $M_1) = 1 - P(M/M$ or $M_1)$, then the posterior analysis is completed by finding the corresponding posterior probabilities $P(M/\mathbf{y}_d$, M or $M_1)$ and its complement. This can be done using Bayes' theorem, which gives:

$$\frac{P(M/\mathbf{y}_d,M \text{ or } M_1)}{P(M_1/\mathbf{y}_d,M \text{ or } M_1)} = F.\ \frac{P(M/M \text{ or } M_1)}{P(M_1/M \text{ or } M_1)}$$

where

$$F = \frac{p(\mathbf{y}_d/M)}{p(\mathbf{y}_d/M_1)}$$

is the so-called Bayes factor, which converts prior odds into posterior odds. This is where the predictive density enters the analysis, but since the approach is Bayesian and obeys the Likelihood Principle, only the predictive density for the observed $\mathbf{y}_d$ is relevant. By looking at tail-area probabilities, involving $p(\mathbf{y}/M)$ for other values of $\mathbf{y}$, Professor Box is making a fundamental departure from the correct Bayesian solution. Why should he do this?

Perhaps the answer is that his approach seems to avoid the need to specify the alternative model M_1. Formally, of course, we cannot discredit M without consideration of alternatives. It is to be discarded if $p(\mathbf{y}_d/M)$is small *not* relative to the value it might have taken had some other sample been observed, but relative to the value it would take under some viable alternative M_1. The word "viable" is to convey the fact that $P(M_1/M$ or $M_1)$ should not be extremely small, otherwise a very small value of F need not lead to posterior odds strongly favouring M_1.

In practice we cannot formally consider all the possible alternatives, and if Professor Box has succeeded in avoiding the need for them then this is quite an achievement. He actually refers to the way his procedure might be applied informally, in practice, as follows.

> "In practice... diagnostic checking... is often conducted by visual inspection of residual displays or other more sophisticated plots... The statistician is looking for features in the data which would be surprising or unusual if the model M were true. Such a feature can be described by a function $g(\mathbf{y}_d)$ and its unusualness... measured by reference to $p(g(\mathbf{y})/M)$."

The reason for suddenly introducing $g(\mathbf{y}_d)$ is mentioned in his preceding paragraph, but is much better shown in an example which unfortunately does not appear in the shortened version of the paper. This example was of a sample, according to M, from a normal distribution. In diagnostic checking in relation to this example, he clearly has in mind the possibility of outliers as one potentially surprising feature of the data. But the predictive density $p(\mathbf{y}_d/M)$ depends only on the sufficient statistics s^2 and $\bar{y}$. Therefore it registers only weakly the surprise we feel when the data suggest the presence of outliers, for then it is more the pattern of data points than their location or spread which catches our eye. But clearly Professor Box can choose a $g(\mathbf{y}_d)$ which would register our surprise much more strongly. This is why $g(\mathbf{y})$ is a necessary artefact in his approach, but of course the choice of $g(\mathbf{y})$ is no different from a choice of alternative model.

The correct Bayesian approach makes it clear that surprise is not enough. What a practising statistican does when he looks for surprising and interesting features in his data is more sophisticated than Professor Box supposes. He may have no alternatives in mind explicitly beforehand, and may find it difficult to formulate one afterwards, but *viable* alternatives are implicit in all the ways in which he chooses to look at his data. This is where his skill and experience tell - in what he chooses to look at, in what he registers surprise at. His reaction signifies not only that $p(\mathbf{y}_d/M)$ is small (surprise!) but also that his experience tells him that he will probably be able to find an alternative M_1 such that $p(\mathbf{y}_d/M_1)$ is much larger, i.e. the surprise is removed, and such that $P(M_1/M$ or $M_1)$ is not negligible.

The case of surprising outliers leads neatly to Professor Freeman's paper. He presents three different alternative models, each of which allows a mechanism for the occurrence of outliers. Each would in general greatly reduce the level of surprise we would feel when confronted by data exhibiting outliers, but each mechanism is different. Consider Professor Freeman's analysis of the Darwin data. On the assumption that there are two outliers the Abraham-Box model fails to identify "the most obvious pair (-67, -48)" as the culprits, and he concludes that "The [AB] model is clearly not a good one for identifying outliers". The conclusion is far too strong. The point is that if we believe the AB model to be appropriate then (-67, -48) is *not* a terribly obvious outlier pair, since to accommodate both these as outliers with a single value of the discrepancy parameter δ still necessitates large residuals. The element of surprise is still quite strong. Whereas under the BT model, for example, the Darwin data would be much less surprising. The conclusion is that *if* the BT model were *a priori* viable then the data would favour it through the Bayes factor F, and we would say that the AB model is probably not correct *for these data.*

Professor Freeman's other examples are similar. What he sees as an outlier may not be the kind of outlier generated typically by one or other of the three models. Performance is inversely related to surprise. The examples are instructive because they tell us something about the different outlier-producing mechanisms of the various models, which in practice will help us to assess prior probabilities.

It is interesting that by focussing his attention on *identifying* outliers Professor Freeman places very different emphasis from Professor Box, who would be more concerned with estimating β. The unstated implication is that all three methods would yield robust inference about β, but this is not true. The AB method simply gives suspected outliers a reduced weight, and if they deviate far enough from the others their influence can be strong. In O'Hagan (1979) I have looked at how robustness can be achieved simply by assuming that the data are sampled from a distribution with a suitably thick tail. Outlier rejection will then take place regardless of our prior distribution for β. It is interesting that, in an earlier paper than their one on outliers, Box and Tiao (1962) examined the Darwin data under thick-tailed alternatives, but that none of their distributions had thick enough tails to guarantee outlier rejection (see O'Hagan (1979)). I hope to publish numerical results soon.

I would like to end by emphasising that I found both papers profoundly stimulating, and that, if I have appeared to be highly critical, this is merely because the questions they raise are so important and so deep. I would like to congratulate and to thank both authors.

J.M. BERNARDO (*Universidad de Valencia*):

Professor Box's thought provoking paper distinguishes between model criticism and parameter estimation and goes on to advocate a (conditional) Bayesian analysis for the latter but a frequentist-type one for the former. I feel that the division between model and prior is somewhat illusory. What one really needs is the joint distribution $p(x,\theta)$ and it is only tradition which gives $p(x|\theta)$ and $p(\theta)$ a different theoretical status. Indeed when one uses some sort of plot to 'test' empirically $p(x|\theta)$ what one is really 'testing' is rather the predictive $p(x)$. Whether you call $p(x,\theta)$ a 'model' or 'a prior' is unimportant, but it seems to me that empirically testable prediction conditional to $p(x,\theta)$ is often what is precisely needed.

P.J. BROWN, (*Imperial College, London*):

Some of the discussion on outliers so far today does seem a little unreal. In my experience identification of an outlier is just a signal to investigate further. On closer inspection and with more data there may well be good reasons to so regard it. In election night forecasting, for example, 'stringers' waiting at the counting halls are relied upon to telephone in the results as soon as they are declared. It is understandable that a few may take to alcohol to while away the long night. An absurd result, if flagged, will result in further corroboratory telephone calls to the constituency. Thus this outlier problem is sequential.

I would like to see much more precision in the definition of the term 'outlier'. Obviously there are workable definitions outside that of data transmission errors but, without more careful examination of the utility of the concept and its realisation, I think one cannot proceed beyond accepting that there are a number of different possible conclusions, each having some plausibility.

A.P. DAWID (*The City University*):

It is not necessarily true, as Professor Box suggests, that the use of improper priors does not allow model criticism. Suppose our observation is y, with the binomial distribution $\mathbf{B}(n;\theta)$, and we use the improper prior distribution $\beta(0,0)$, viz. $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, considered, say, as a limit of $\beta(\alpha,\alpha)$ with $\alpha \to 0$. The limiting predictive distribution has $P(y = 0) = P(y = n) = ½$, so that any value $0 < y < n$ discredits this "model-cum-prior". However, if we believe the weak spot to be in the prior specification, rather than the sampling model, we should not be too hasty to discard our assumptions, since our posterior distribution is not likely to be sensitive to the choice of prior. Somewhat paradoxically, it is for the case $y=0$ (or n), which does *not* discredit the improper prior, that we must be most careful about specifying the "true" prior distribution. This example indicates to me that model-checking using predictive distributions may not always be appropriate.

J.M. DICKEY (*University College Wales, Aberystwyth*):

Professor Freeman has not made the assumption of condition continuity in his paper here on outliers, in the sense that in Section 4 his prior opinion concerning a single outlier is not necessarily the same as if he had been told that of two outliers one had zero disturbance. I am wondering what kind of relationship one would want between these integrable prior distributions conditional on different models. (See my discussion to the paper by Professor A.F.M. Smith in these Proceedings).

I don't like the assumption in Section 1 of a uniform improper prior distribution in as many dimensions as the number of outliers (G.D.F. approach). In principle, the number of dimensions can be as high as the sample size, and constant nonintegrable densities are notoriously troublesome in high dimensions.

I hope Professor Freeman will develop further his interesting integrable-prior methods (Section 4), and report his experience in their use bearing on the important questions of choice of prior distribution.

I.J. GOOD (*Virginia Polytechnic and State University*):

I am pleased to see that so distinguished a statistician as Professor Box has emphasized a Bayes/non-Bayes compromise or synthesis because that has been my philosophical position for a long time, although I regard the Bayesian side of it as more fundamental. One example of such a marriage, especially close to the theme of Professor Box's lecture, is the use of orthodox significance tests for choosing a hyperparameter, and for testing a Bayesian model, for density estimation and bump-hunting. This idea was presented in August 1974 in the invited General Methodology lecture at the annual meetings of the American Statistical Association in St. Louis, Missouri. Practical applications of the method are given in Good and Gaskins, (1980). In the Journal of the American Statistical Association, **75** (1980), 42-73 (with discussion).

By saying that the Bayesian side of the coin is more fundamental I mean that the use of tail-area probabilities can be roughly justified by Bayesian arguments when it can be justified at all. (See my contribution to Professor Barnard's seminar for references).

A.F.M. SMITH (*University of Nottingham*):

Box argues that *criticism* must ultimately appeal to sampling theory for its justification. He may well be correct, but I am not convinced that the development given here succeeds in clearly demarcating an area of critical activity that is out of bounds to a Bayesian. There would seem to be, in broad terms, a one-to-one relationship between any diagnostic checking procedure and an *implicit* family of alternative models. Indeed, Box comes close to conceding the primacy of such implicit alternatives when he turns to "Choosing the diagnostic checks". The ensuing discussion of "Diagnostic checking and Robustification" appears to acknowledge this one-to-one correspondence and thus, surely, to admit that whatever can be probed using a diagnostic check function can also be probed by using Bayes factors against appropriate alternative models. Some of the author's general discussion seems intended as a defence against this latter accusation, but it has equal force, or rather lack of it, against *both* approaches. Either we attempt no criticism (i.e. *no* diagnostic checks, *no* Bayes factors) or we attempt *some* limited criticism (i.e. apply a *finite* number of diagnostic checks, calculate a corresponding *finite* number of Bayes factors). In neither case can we test against all possible departures (using *all possible* diagnostic checks, or a *totally comprehensive* model).

I am not disposed to think that "it" (the advancement of learning?) can all be done with Bayes, but I do feel that the kinds of *local* model criticism discussed in this paper *can* be carried out within the Bayesian framework and that, at most, we are here discussing rather pragmatic issues and not fundamental questions about inferential paradigms.

REPLY TO THE DISCUSSION

P.R. FREEMAN (*Leicester University*):

Several discussants mentioned the need for a proper definition of an outlier, so that we are all clear what we are talking about. It seems to me impossible to ever get a fully operational definition, although we can all recognise an outlier when we see one, since if we try to model formally all the possible kinds of outlier, we shall end up with something which is far too complex to be of any use. For example, Professor Eddy's suggested model gains in flexibility, certainly, but loses in complexity since we would have to take a double sum over all values of r_1 and r_2, and the combinatorial explosion would defeat us for even very small sample sizes.

I think that outlier identification is important since ideally we want to do the sequential checks just as Dr. Brown describes (and to ensure that the faulty "stringers" are not employed at the next election). There is no real substitute for the hard work of going back over records and finding the exact source of error (or for failing to find any error), and for then re-analysing the data with the suspicious values either corrected, deleted or left unchanged. But in the real world this is just far too much trouble and some robustness of analysis is also desirable so as to save much of this work. It was in this sense that I criticised the AB model. I should have said that it is not flexible enough to detect some kinds of outliers that I think I would like to have detected, namely those occurring at both ends of the data.

I take Dr. O'Hagan's point that we need some automatic protection against very extreme observations. The GDF model does this by ignoring them completely, but I agree that models with thick tails should be used in many situations where we dangerously use normal tails at present. Dr. Eddy finds this aspect of GDF unattractive, but I would justify it by saying that the overall effect is somewhat comparable to that of the jack knife with the more sensible refinement of taking a weighted average of the results obtained by dropping one or more observations at a time. The extremely deviant values only get ignored completely when they are so far out that one subset attracts all the posterior weight to itself.

I thank Dr. Eddy for unifying the notation of the 3 models. I only wish I had thought of doing so when I wrote the paper.

Professor Dickey comes close to the heart of the problematic area of my paper-the choice of priors. I, too, am perturbed by the improper priors in the GDF model, though they do in practice give beautifully robust results for parameter estimation. I am not too worried by the lack of condition continuity in my priors as I can see no intuitively compelling reason to obey that condition and it is not, as far as I can tell, an essential requirement for coherence. The dependence on the exact form of the conditioning again makes me sceptical of its usefulness.

The proper-priors section of my paper still seems to me to contravene what was enunciated verbally at the conference as Lindley's principle - that if you take a problem, treat it coherently and use sensible priors you will always get a sensible answer. It is not clear to me what part of the conditions I am violating, thought the answers I get are disappointingly misleading. Perhaps the attemps to discriminate among members of a nested family of hypotheses is doomed to failure due to lack of enough data, whatever the priors. Only further work and deeper consideration will tell.

G.E.P. BOX (*University of Wisconsin*):

It is perhaps hardly surprising that I have not been totally successful in convincing a conference of Bayesians of the auxilliary need for Sampling Theory and I have sympathy with some of my critics.

In response to Professors Smith, O'Hagan and Eddy, my main point is that since Bayes is conditional, if it is to be used exclusively in the pursuit of an adequate model, we inevitably find ourselves engaged in a game of "Yes but". It is rather as if, when I was preparing for my early morning dash to the airport on leaving Los Fuentes, my conversation with the hotel manager had gone as follows:

> Do you think I can catch my plane?
> Yes, if the taxi is on time.
> Do you think the taxi will be on time?
> Yes if the taximan gets up early enough.
> Do you think he will get up early enough?
> Yes if his wife remembers to wake him.
> etc., etc.

More specifically, *however* far the model building process had been carried by Bayesian methods the final model would still be

$$p(\mathbf{y},\theta\,|\,M_k) = p(\theta\,|\,\mathbf{y},M_k)p(\mathbf{y}\,|\,M_k)$$

and there remains the n-dimensional space of the marginal predictive distribution $p(\mathbf{y}\,|\,M_k)$ which has not yet been explored and which can, on a sampling theory argument, discredit the relevance of the assumptions on which the Bayesian analysis is conditional.

I grant that, as soon as we start to consider specific alternative models, then

Bayesian versions of diagnostic checks are available. In particular for the case of a discrepancy parameter β taking the value $\beta = \beta_0$ for an ideal model M, one way in which this duality may be formalized is as follows. A natural function of the data to consider for making diagnostic checks is

$$g_\beta(\mathbf{y}) = \left.\frac{\partial \log p(\mathbf{y}|\beta)}{\partial\beta}\right|_{\beta = \beta_0}$$

But since $p_u(\beta|\mathbf{y}) = p(\beta|\mathbf{y})/p(\beta)$ we see that $p_u(\beta|\mathbf{y}) = p(\mathbf{y}|\beta)$ so that $g_\beta(\mathbf{y})$ is Fisher's score function for the parameter β. So it may be argued why not just look at the distribution $p_u(\beta|\mathbf{y})$?

The amount of effort that can be expended on any particular analysis is finite and we may not want to expend a full Bayesian analysis on every discrepancy that occurs to us. In many cases the model builder would be satisfied with graphical checks. Even so such checks need not be entirely ad hoc and indeed it is possible to show that $g_\beta(\mathbf{y})$ defined above is often valuable in showing the form that graphical checks should take.

I, of course, agree with Dr. O'Hagan that the predictive ratio $p(\mathbf{y}|M_1)/p(\mathbf{y}|M_0)$ can be used not only to indicate the appropriate form for diagnostic checking functions, but also in the direct Bayesian assessment of the relative evidence of any one model versus another. Notice, however, that the inherent Bayesian limitation of conditionality ensures that, however large this ratio may be, the preferred model M_1 can still be manifestly implausible because $Pr\{p(\mathbf{y}|M_1) < p(\mathbf{y}_d|M_1)\}$ is small.

I am grateful to Professor Good for his encouraging comments and references.

Consider Professor Dawid's example when the limit $\alpha = 0$ is *not* approached, remembering to make due allowance for the fact that while θ is continuous y is discrete. The choice of prior $\beta(\alpha,\alpha)$ is equivalent to supposing a uniform prior in $\phi = \int^\theta t(1-t)^{\alpha-1}dt$. If we take $\alpha = 1$ the predictive distribution $p(y|M)$ is such that $p(y/N|M) = (N+1)^{-1}$, $(y = 0,1,2,\ldots N)$ and the predictive cumulative distribution plots as a linear "staircase function" against y/N. Thus supposed indifference about θ itself results in no predictive critical ability for y/N. But suppose following Jeffreys we set $\alpha = ½$, then $\phi = \sin^{-1}\sqrt{\theta}$, $0 \leq \phi \leq \pi/2$. The corresponding predictive distribution for $\sin^{-1}\sqrt{(y/N)}$ is, of course, unequally spaced but again the cumulative distribution even for small samples approximates a straight line and supposed indifference about $\sin^{-1}\sqrt{\theta}$ results in no predictive critical ability for $\sin^{-1}\sqrt{(y/N)}$. The approximation holds for other non-zero values of α, however, as we go to the limit $\alpha = 0$ the range for ϕ goes from $-\infty$ to $+\infty$ and consequently the discrete predictive distribution is dominated by values corresponding to $y = 0$ and $y = N$ which are infinitely removed from other realizations. I would argue, therefore, that this example reconfirms the unsuitable nature of this particular prior, the unsuitability of which as Professor Dawid says is not clear from consideration of the posterior distribution which over the range considered is sensitive to the changes discussed. In choosing prior distributions we must clearly consider their predictive consequences.

Although I much enjoyed this Bayesian Conference, there was for me an eerie feeling that something important was missing. Bayesian inference is an instrument for use in scientific enquiry. But except for a couple of rather distant echos we seemed to have talked for a week securely insulated from the world of real investigation. It has been said that

> "Theory and Practice are like man and wife in a happy marriage; each complements and inspires the other and without interaction between them there can be no new life".

Certainly the work of such practicioners as Gauss, Laplace, Daniel Bernoulli, Fisher and Jeffreys provides no reason to doubt this aphorism.

I believe it is agreed that scientific iteration employs in alternation the dual processes of model criticism on the one hand and exploitation of the tested model on the other. Suppose we accepted, as I suggested in my paper, that two different kinds of inference are needed to conduct these two different activities conveniently. Suppose it was agreed that the first activity (which subsumes model specification/identification and tests of fit) although often conducted informally under the name of Exploratory Data Analysis ultimately requires Sampling Theory for its justification, while the second requires Bayesian Theory. Then it would be understandable why a purely Bayesian conference would have little to say about any real scientific investigation (and perhaps a conference entirely devoted to "Exploratory Data Analysis" might be equally disappointing).

It is rather as if we called a conference of airplane pilots* who knew everything about landing a plane but nothing about how to take off (or vice versa). At such a conference there should be little surprise if in a welter of papers viewing from every angle the finer theoretical points of landing an airplane the discussion seldom turned on going anywhere or on interesting voyages experienced.

* They might more properly be called "landers" rather than pilots, just as some of us are called Bayesians rather than Statisticians.

REFERENCES IN THE DISCUSSION

BOX, G.E.P. and TIAO, G.C. (1962). A further look at robustness via Bayes's theorem. *Biometrika* **49**, 419-432.

GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scatering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 42-73 (with discussion).

O'HAGAN, A. (1979). On outlier rejection phenomena in Bayes inference. *J. Roy. Statist. Soc. B*. **41**, 358-367.

Trab. Estadist (1980)
Vol. 31, pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 383-430 Valencia, Spain: University Press.

9. Improving judgements using feedback

INVITED PAPERS

DE GROOT, M.H. (*Carnegie-Mellon University*)
Improving predictive distributions

PRESS, S.J. (*University of California, Riverside*)
Bayesian inference in group judgement formulation and decision making using qualitative controlled feedback

DISCUSSANTS

DUNSMORE, I.R. (*University of Sheffield*)
GEISSER, S. (*University of Minnesota*)
BERNARDO, J.M. (*Universidad de Valencia*)
DAWID, A.P. (*The City University*)
DUMOUCHEL, W.H. (*Massachussets Institute of Technology*)
FRENCH, S. (*University of Manchester*)
GOOD, I.J. (*Virginia Polytechnic and State University*)
LINDLEY, D.V. (*University College, London*)
ZELLNER, A. (*University of Chicago*)

REPLY TO THE DISCUSSION

Improving Predictive Distributions

M.H. DeGROOT
Carnegie-Mellon University

SUMMARY

Consider a sequence of decision problems S_1,S_2, ... and suppose that in problem S_i the statistician must specify his predictive distribution F_i for some random variable X_i and then make a decision based on that distribution. For example, X_i might be the return on some particular investment and the statistician must decide whether or not to make that investment. The random variables X_1,X_2, ... are assumed to be independent and completely unrelated. It is also assumed that each predictive distribution F_i assigned by the statistician is a subjective distribution based on his information and beliefs about X_i. In this context, the standard Bayesian approach provides no basis for evaluating whether the statistician's subjective predictive distribution for X_i is good or bad, and does not even recognize this question as being meaningful. In this paper we describe models in which the statistician can study his process for specifying predictive distributions, identify bad habits, and improve his predictions and decisions by gradually breaking these habits.

Keywords: SUBJECTIVE PROBABILITY; PRIOR DISTRIBUTIONS; SPECIFICATION BIAS; LEARNING; SEQUENTIAL DECISION MAKING; PREDICTIVE DISTRIBUTIONS.

1. INTRODUCTION

Consider a statistician who must specify his subjective predictive distribution for some variable X. Suppose that after duly considering all of his available information, the statistician specifies a particular normal distribution as his predictive distribution for X. Suppose also that the value of X is subsequently observed and is found to be far out in the tail of the statistician's distribution.

How should the statistician react to this observation? The Bayesian approach to statistics provides no answer to this question. In fact, it does not even recognize the question as being relevant or meaningful. As long as the statistician supplies a proper probability distribution as his predictive distribution and makes decisions that are optimal with respect to that distribution, and as long as the observed value of X actually lies in the support of his predictive distribution, there is no reason to question that distribution and no mechanism for doing so.

Although much has been written on how to specify a prior or predictive distribution [see e.g., Savage (1971) and Hogarth (1975)] and on the pitfalls for the unwary statistician [see, e.g., Spetzler and Staël von Holstein (1975) and Tversky and Kahneman (1974)]; little or nothing has been written on how to check back to see if the distribution was reasonable. Indeed, a prior or predictive distribution is regarded as the statistician's personal opinion and it is not to be challenged or questioned.

In practice, however, I believe that there are four possible reactions to an observation that falls far out in the tail of a predictive distribution: (1) The statistician could feel that his prior information was bad and misleading, in the sense that it had indicated that X was not likely to lie in the region in which it actually did lie. (2) He could feel that a rare event had occurred because a very unusual observation has been obtained. (3) He could feel that he had made his predictive distribution more concentrated than he should have solely on the basis of his available information, and that his distribution should have been more spread out in order to better accommodate the observed value of X. (4) He could feel that there is not, and should not be, any relation between his predictive distribution and the observed value.

We can probably rule out reaction (4) on the pragmatic basis that the statistician should try to make his predictive distribution represent his view of where X is likely to fall as much as possible. However, it is typically not possible to decide which of the first three reactions is most appropriate on the basis of just a single observation unless two or more statisticians have specified their predictive distributions for X, in which case we can make comparisons among them [see e.g., Roberts (1965) and DeGroot (1970), page 154]. In this paper we shall study just one statistician who must repeatedly specify his predictive distributions in many different problems as they arise.

Specifically, we shall consider a sequence of similar but unrelated decision problems S_1, S_2, ... and we shall suppose that in problem S_i the statistician must specify his predictive distribution F_i for some random variable X_i and possibly make a decision based on that specification. For example, X_i might be the return on some particular investment, and the statistician may have to decide whether or not to make that investment. The statement that the problems S_1,S_2, ... are unrelated is meant to mean the same thing as the assumption that the variables X_1,X_2, ... are independent.

Suppose that the statistician finds that in a large proportion of these problems, say 80 percent of them, the observed values of the variables lie far out in lower tail of the specified predictive distributions, say more than five standard deviations from the mean. Since Bayesian theory tells the statistician that his predictive distributions are just as valid as anyone else's, he might simply think that he has had an unusually severe run of bad luck. It is much more reasonable, however, for the statistician to feel that in some sense the world is different from his perception of it as represented by his predictive distributions, that his perception has some systematic bias, and that he should change his perception and his predictive distributions. This idea is also mentioned by Winkler (1967).

A "rational" person cannot go through life continually being surprised by his experiences and observations. After a while, he will revise his perception so that the unexpected becomes the expected, and observations that were formerly surprising become routine. The "rational" person who regards the rising of the sun each morning as a sequence of independent trials may be surprised for a while to discover the sun each morning at dawn, but after a week or two he will merely yawn, turn over in bed and go back to sleep.

In this paper we shall present some models of the process by which a statistician adjusts his perceptions and the way in which he specifies predictive distributions. In effect, these are models of how the statistician can learn about his own biases and errors in the specification of predictive distributions, and how he can adjust for them and gradually eliminate them.

Before beginning the development of these models, we conclude this section with two comments:

(1) Since we are studying a sequence S_1, S_2, ... of similar and independent decision problems, we could model the process by which the statistician adjusts his predictive distribution in "frequency" or "sampling theory" terms. For example, we could develop models in which statistician changes the way he specifies predictive distributions if some test of significance indicates that the observed X_i's were not generated by the predictive distributions F_i. However, since we are trying to model the process by which a Bayesian statistician adjusts his subjective distributions, it seems more apt to use a Bayesian model.

(2) Although the decision problems S_1,S_2, ... are independent, they do have one element in common that permits learning and adaptation from earlier S_i's to later ones. The common element is the statistician himself, and it is the statistician's behavior that we are trying to model.

2. NORMAL PREDICTIVE DISTRIBUTIONS

We shall continue to consider a sequence of decision problems S_1, S_2,... in which the statistician must specify his predictive distributions for X_1,X_2, ... It will still be helpful to think of the specific context in which X_i is the return on some investment, and we shall use that terminology whenever it is convenient. Throughout the remainder of this paper we shall assume that although the variables $X_1, X_2, \ldots$ are independent, the investment problems in the sequence are similar in nature in the sense that the statistician has similar types of information in each problem and it is natural for him to specify a normal predictive distribution for each X_i.

We shall use the notation $Z \sim N(m,r)$ to indicate that a random variable Z has a normal distribution with mean m and precision r. (The precision is the reciprocal of the variance.) If the statistician specifies that his predictive distribution for some random variable X is $N(m,r)$ then his predictive distribution for $Y = r^{1/2}(X-m)$ will be the standard normal distribution $N(0,1)$.

Rather than specifying his normal predictive distribution for X, the statistician could equivalently specify the particular linear transform Y for which his predictive distribution is $N(0,1)$. Thus, we shall assume that in each decision problem S_i, the statistician characterizes his predictive distribution for the return by specifying a random variable Y_i that is some linear transform of X_i for which the predictive distribution is $N(0,1)$.

Suppose now that in the first problem S_1, the statistician is just about to specify that $Y_1 \sim N(0,1)$ when he remembers, based on the returns from previous investments under similar conditions, that he may have a tendency to specify predictive distributions that are shifted too far to one side or the other. In other words, the statistician now recognizes that because of his tendency to misspecify the mean of his normal predictive distributions, he should actually specify that $Y_1 \sim N(\theta_1,1)$.

However, the value of θ_1, which represents the statistician's bias, is not known to him. If $\theta_1 > 0$, the statistician is a pessimist in the sense that his predictive distribution tends to underestimate the actual return. If $\theta_1 < 0$, he is an optimist in the sense that his predictive distribution tends to overestimate the actual return.

Since the value of θ_1 is not known to the statistician, he must assign a prior distribution to it which reflects his own beliefs about whether he is likely to be a pessimist or an optimist. We shall assume that this prior distribution is again a normal distribution $N(\mu_1,\tau_1)$. If $\mu_1 > 0$ ($\mu_1 < 0$) the statistician feels that he is probably a pessimist (an optimist).

The assignment of this prior distribution to θ_1 has two effects. First, the statistician must change his predictive distribution for Y_1. Since the conditional distribution of Y_1 given θ_1 is $N(\theta_1,1)$ and $\theta_1 \sim N(\mu_1,\tau_1)$ it follows that the marginal or predictive distribution of Y_1 is $N[\ \mu_1,(\tau_1/\tau_1+1)]$. Second, the statistician can learn about his specification bias θ_1 by observing the value of Y_1.

The posterior distribution of θ_1 given that $Y_1 = y_1$ is $N(\mu_1', \tau_1')$, where

$$\mu_1' = \frac{\tau_1\mu_1 + y_1}{\tau_1 + 1} \quad \text{and} \quad \tau_1' = \tau_1 + 1. \tag{2.1}$$

The statistician can study this posterior distribution to learn what kind of specification bias he is likely to have made in his initial standard normal predictive distribution for Y_1.

It should be emphasized that in all the models to be considered in this paper, learning always proceeds in accordance with Bayes' theorem. Models in which there is non-Bayesian learning have been discussed by Lad (1978) and others.

3. REDUCTION OF THE SPECIFICATION BIAS

When the statistician now moves on and faces the decision problem S_2, he must again specify a random variable Y_2 for which his predictive distribution is $N(0,1)$. Again, he will recognize that the Y_2 that he would like to select is subject to a specification bias and that he should actually specify that the predictive distribution of Y_2 is $N(\theta_2,1)$. We now come to a crucial step. What prior distribution for θ_2 should the statistician use at this stage? Should he use the posterior distribution of θ_1 obtained from the problem S_1? In other words, will his specification bias θ_2 in the problem S_2 be the same as his specification bias in S_1?

It would be very tempting to assume that the answer to these last two questions was "yes", for we could then proceed in a completely straightforward fashion. The specification bias θ_1 would remain fixed throught the entire process, and as the statistician moved on thorought the sequence of problems $S_1, S_2, \ldots$ he would learn more and more about its value until he knew it almost exactly.

But consider what this model would imply about the behavior of the statistician. It would condemn him to a life of second-guessing himself. It would imply that in each problem that he faced, he would first feel that his predictive distribution for a certain variable Y_i was $N(0,1)$, he would then remember his specification bias θ_i, and he would then second-guess himself and specify an adjusted predictive distribution for Y_i.

I do not believe that such behavior is reasonable, or even stable, in the long run. If the statistician learns that he must always second-guess his own predictive distributions, then he will begin to anticipate this aspect in his initial specification, and soon he will be "third-guessing" himself, then "fourth-guessing" himself, etc. For this reason, we shall not follow this approach here.

I believe that the prior distribution of θ_2 will not be the same as the posterior distribution of θ_1. The statistician will learn something from observing the outcome of S_1 about the type of specification errors that he is likely to make and, consciously or subconsciously, he will use this knowledge about his own personality when he initially specifies the predictive distribution of Y_2. Hence, the specification bias for Y_2 is likely to be smaller that it was for Y_1.

The main thrust of the remainder of the paper will be to develop models in which the statistician tends to reduce his specification bias in succesive problems and ultimately eliminates it entirely. Thus, in the limit, the predictive normal distribution that he initially specifies at the beginning of a decision problem will actually represent his subjective opinion without any further second-guessing or other adjustments. We shall now present some simple models of this type.

4. MODELS OF THE LEARNING PROCESS

After the value $Y_1 = y_1$ has been observed in the decision problem S_1, the statistician will calculate the posterior mean μ_1' of his specification bias θ_1 and carefully contemplate its value. Since μ_1' is the statistician's mean value for his specification bias, we shall assume that the result of his careful contemplation is that he subconsciously changes his bias in the next problem by an amount equal to some fraction of μ_1'. In brief, we shall assume that

$$\theta_2 = \theta_1 - \gamma\mu_1', \tag{4.1}$$

where γ is a fixed number $(0<\gamma\leq 1)$.

The entire process evolves as follows: In the problem $S_n(n=1,2,\ldots)$, the statistician will initially be tempted to specify that the predictive distribution of a certain variable Y_n is $N(0,1)$. He will recognize, however, that because of his specification bias, he should actually specify that the predictive distribution of Y_n is $N(\theta_n,1)$. The statistician will assign a prior distribution to θ_n and, after observing the value $Y_n = y_n$, he will calculate the posterior distribution of θ_n.

We shall assume that a relation like (4.1) holds throughout the process. Thus, we assume that

$$\theta_{n+1} = \theta_n - \gamma\mu_n' \qquad (n = 1,2,\ldots) \tag{4.2}$$

where μ_n' is the posterior mean of θ_n.

We have already assumed that the prior distribution of θ_1 is $N(\mu_1, \tau_1)$. It now follows that both the prior and posterior distsributions of every θ_n will be normal. If the prior distribution of θ_n after $Y_1,\ldots,Y_{n-1}$ but not Y_n have been observed is $N(\mu_n,\tau_n)$ and the posterior distribution after Y_n has been observed is $N(\mu_n', \tau_n')$ then the following relations are statisfied:

$$\mu_{n+1} = (1\text{-}\gamma)\mu_n' = (1\text{-}\gamma)\ \frac{\tau_n\ \mu_n + y_n}{\tau_n + 1} \tag{4.3}$$

$$\tau_{n+1} = \tau_n' = \tau_n + 1 = \tau_1 + n. \tag{4.4}$$

It follows from (4.3) and (4.4) by induction that, for $n = 1,2,\ldots$,

$$\mu_{n+1} = \frac{1}{\tau_1 + n}\,[(1\text{-}\gamma)^n\,\tau_1\mu_1 + \Sigma_{j=1}^n (1\text{-}\gamma)^{n+1-j}\,y_j]. \tag{4.5}$$

As discussed in Section 3, we are interested in conditions under which $\theta_n \to 0$ in some appropriate sense. Since $\theta_n \sim N(\mu_n,\tau_n)$, and $\tau_n \to \infty$ by (4.4), it follows that $\theta_n \to 0$ in probability if $\mu_n \to 0$.

If $\gamma = 1$, then it can immediately be seen from (4.3) that $\mu_{n+1} = 0$ for $n = 1,2,\ldots$ Hence, if the statistician's learning process as represented by (4.2) is such that the mean of his specification bias is 0 at each stage after the first, then the statistician will ultimately eliminate his specification bias.

Now suppose that $0 < \gamma < 1$. Then, from (4.5), $\mu_n \to 0$ if and only if

$$\lim_{n\to\infty} 1/n\,\Sigma_{j=1}^n (1\text{-}\gamma)^{n+1-j}\,y_i = 0. \tag{4.6}$$

The relation (4.6) will be satisfied for most sequences of observations y_1, y_2, ... In fact, (4.6) will be satisfied unless $|y_n|$ grows large at a relatively fast rate. Thus, the statistician will ultimately eliminate his specification bias *unless there is something about the specification and learning process that* leads the random variables $Y_1, Y_2, \ldots,$ to fall farther and farther out into the tails of their predictive distributions. In particular, it can be shown that the specification bias will be eliminated if $A_n = 1/n\,\Sigma_{j=1}^n |y_j|$ converges to a finite limit or more generally, if A_n is bounded as $n \to \infty$.

5. LEARNING ABOUT LEARNING

The constant γ that appears in Eq. (4.2) characterizes the learning process of the statistician. Presumably, different statisticians would subconsciously reduce their specification bias at different rates, and they would therefore have different values of γ. In this sense, γ can be regarded as another parameter of the model.

It is not necessary for the statistician to assume that γ simply has a particular value. He can assign a prior distribution to γ at the beginning of problem S_2 and learn about its value as the process evolves. With this approach, for $n = 2,3,\ldots,$ the conditional distribution of Y_n given θ_n and γ will be $N(\theta_n,1)$. The conditional prior distribution of θ_n given γ, based on the observed values of $Y_1,\ldots, Y_{n-1}$ but not of Y_n, will be $N[(1-\gamma)\mu'_{n-1},\tau'_{n-1}]$, just as before, and γ will have some specified marginal prior distribution ζ_n on the interval $0 < \gamma \leq 1$.

Since the statistician is now uncertain about the values of both θ_n and γ, the actual predictive distribution of Y_n that he must specify will be the marginal distribution of Y_n obtained by integration over the joint prior distribution of θ_n and γ. After Y_n has been observed, the posterior conditional distribution of θ_n given γ will be $N(\mu'_n, \tau'_n)$, just as before, and the posterior distribution of γ will be some new distribution ζ'_n. The prior distribution of γ for the problem S_{n+1} will be $\zeta_{n+1} = \zeta'_n$.

There do not seem to be any natural or neat prior distributions for γ, and we shall not pursue any calculations here. Until the statistician has learned his value of γ, he will be forced to specify nonnormal predictive distributions for Y_n.

The important point of this discussion is that it is possible for the statistician to learn about his learning rate. Of course, this idea could be further developed in a hierarchical model in which a prior distribution is assigned to hyperparameters that appear in the prior distribution of γ. In this way, the statistician can learn about the rate at which he learns about his learning rate, etc. We shall not explore this topic in this paper.

6. MISSPECIFICATION OF THE PRECISION

Suppose now that instead of the statistician tending to specify predictive distributions that are shifted too far to the left or the right, he tends to specify predictive distributions that are appropriately centered but are either too concentrated or too widely spread out. We shall assume that when the statistician is tempted to state that his predictive distribution for a certain variable Y_n is $N(0,1)$, he recognizes that because of his tendency to misspecify the distribution he should actually specify the distribution of Y_n to be $N(0,R_n)$.

The statistician does not know the value of R_n. If $R_n > 1$, the interpretation is that the statistician tends to be overly conservative in his predictive distribution and makes it more spread out than he has to. If $R_n < 1$, the interpretation is that he tends to make his predictive distribution more concentrated than is warranted on the basis of his knowledge and information.

The gamma distribution with parameters α and β ($\alpha > 0$, $\beta > 0$), denoted $G(\alpha,\beta)$, is defined by the *pdf*:

$$g(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x>0. \tag{6.1}$$

We shall assume that the prior distribution of R_n, after $Y_1,\ldots,Y_{n-1}$ but not Y_n have been observed is a gamma distribution with parameters α_n and β_n.

Since the conditional distribution of Y_n given R_n is $N(0,R_n)$ and $R_n \sim G(\alpha_n,\beta_n)$, it follows that the marginal *pdf* of Y_n is of the form

$$f_n(y) \propto [1 + \frac{1}{2\beta_n} y^2)]^{-(\alpha_n + 1/2)} \text{ for } -\infty < y < \infty. \tag{6.2}$$

Thus, $(\alpha_n/\beta_n)^{1/2}\, Y_n$ has a t distribution with $2\alpha_n$ degrees of freedom. Because of the statistician's uncertainty about the value of R_n, he must specify this marginal distribution of Y_n as his predictive distribution. The posterior distribution of R_n, after $Y_n = y_n$ has been observed, is $G(\alpha'_n,\beta'_n)$ where

$$\alpha'_n = \alpha_n + \tfrac{1}{2} \text{ and } \beta'_n = \beta_n + \tfrac{1}{2} y_n^2. \tag{6.3}$$

The mean of the distribution $G(\alpha,\beta)$ is α/β and the variance is α/β^2. Therefore, generally speaking, the larger the values of α_n and β_n are and the closer they are to each other, the more concentrated that the distribution of R_n will be around the value $R_n = 1$.

As before, the statistician will study the posterior distribution of R_n in order to learn about, and reduce or eliminate, his specification bias. We shall assume that the learning process is characterized by the following relations for $n = 1,2,\ldots$:

$$\alpha_{n+1} = \alpha'_n + \delta \quad \text{and } \beta_{n+1} = \beta'_n + \delta \tag{6.4}$$

where δ is a given positive constant.

If the statistician is successfully reducing his specification bias as the process evolves, then the distribution of R_n should be getting more concentrated around the value 1 as n increase. It can be seen from (6.4) that the prior mean α_{n+1}/β_{n+1} of R_{n+1} will be closer to 1 than the posterior mean α'_n/β'_n of R_n, and the prior coefficient of variation $\alpha_{n+1}{}^{-1/2}$ of R_{n+1} will be smaller than the posterior coefficient of variation $\alpha_n'^{-1/2}$ of R_n. In this sense, the simple model (6.4) does represent learning by the statistician and reduction of his bias at each stage of the process.

It follows from (6.4) that

$$\alpha_{n+1} = \alpha'_n + \delta = \alpha_n + \delta + \tfrac{1}{2} = \alpha_1 + n(\delta + \tfrac{1}{2}),$$

$$\beta_{n+1} = \beta_n' + \delta = \beta_n + \delta + \tfrac{1}{2} y_n^2 = \beta_1 + n\delta + \tfrac{1}{2} \Sigma_{j=1}^n y_j^2. \tag{6.5}$$

If $E(R_n) = \alpha_n/\beta_n \to 1$ and $\mathrm{Var}(R_n) = \alpha_n/\beta_n^2 \to 0$, then $R_n \to 1$ in probability and the statistician ultimately will completely eliminate his specification bias. It follows from (6.5) that $\alpha_n/\beta_n^2 \to 0$. Furthermore, $\alpha_n/\beta_n \to 1$ if and only if

$$\lim_{n\to\infty} 1/n\, \Sigma_{j=1}^n y_j^2 = 1. \tag{6.6}$$

If the random variables $Y_1, Y_2, \ldots,$ are independent and each actually has a $N(0,1)$ distribution, then Eq. (6.6) will be satisfied with probability 1. More generally, if the learning process is working properly, then the "actual" distribution of Y_n should be approaching a $N(0,1)$ distribution as $n \to \infty$, because the statistician feels that $Y_n \sim N(0,1)$ and his specification bias should be vanishing. There is, however, a fundamental difficulty in speaking about the "actual" distribution of Y_n: In what sense does Y_n have an "actual" distribution, or any distribution other than the subjective probability distribution of the statistician?

Thus, we would hope that (6.6) holds and expect it to hold, but about all we can say with authority is that (6.6) will hold if the statistician's predictive distributions are becoming free of specification bias, and these distributions are becoming free of specification bias if (6.6) holds. That statement may not be very reassuring because of its circularity, but it would be disastrous if it were not true.

As before, the statistician need not assume that the parameter δ has a particular value. Rather, he can assign a prior distribution to δ and actually learn about the value of δ that is appropriate for his own learning rate.

7. CONCLUDING REMARKS

In this paper, we have considered some models in which it is assumed that when a statistician specifies a normal predictive distribution for some variable, there is a bias either in the mean or in the precision of the distribution but not in both. We could obviously present other models in which there is a specification bias in both the mean and precision, so that when the statistician specifies that $Y_n \sim N(0,1)$, he should really be specifying that $Y_n \sim N(\theta_n, R_n)$.

However, we shall not present any models of this type here. It is clear that there is a wide variety of models of the learning process that could be postulated for this two-parameter problem as well as for the more limited one-parameter problems that we have discussed in the paper. Some of these models are similar to the ones that were presented and some are quite different.

At present, we have no substantive basis for choosing among these models. The ones that we have presented here were chosen more or less arbitrarily because of their simplicity, merely to serve as vehicles for conveying the ideas that were to be communicated. It would be worthwhile, I believe, to study actual psychological learning processes in order to be able to build more appropriate models.

ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation under grant SES - 7906386. I am indebted to P.K. Goel and J.B. Kadane for several helpful discussions.

REFERENCES

DEGROOT, M.H. (1970). *Optimal Statistical Decisions*, New York: McGraw-Hill.

HOGARTH, R.M. (1975). Cognitive processes and the assessment of subjective probability distributions, *J. Amer. Statist. Assoc.* **70**, 271-289.

LAD, F. (1978). Embedding Bayes' theorem in general learning rules: Connections between idealized behavior and empirical research on learning. *Br. J. Math. Statist. Psychology*. **31**, 113-125.

ROBERTS, H.V. (1965). Probabilistic prediction. *J. Amer. Statist. Assoc.* **65**, 50-62.

SAVAGE, L.J. (1971). Elicitation of personal probabilities and expectations, *J. Amer. Statist. Assoc.* **66**, 783-801.

SPETZLER, C.S. and STAËL VON HOLSTEIN, C.S. (1975). Probability encoding in decision analysis, *Management Sci.* **22**, 340-358.

TVERSKY, A., and KAHNEMAN, D. (1974). Judgement under uncertainty: heuristics and biases. *Science* **185**, 1124-1131.

WINKLER, R.L. (1967). The quantification of judgement: Some methodological suggestions. *J. Amer. Statist. Assoc.* **62**, 1105-1120.

Bayesian inference in group judgment formulation and decision making using qualitative controlled feedback

S.J. PRESS
University of California, Riverside

SUMMARY

This paper considers the problem of making statistical inferences about group judgments and group decisions using Qualitative Controlled Feedback, from the Bayesian point of view. The qualitative controlled feedback procedure was first introduced by Press (1978), for a single question of interest. The procedure is first reviewed here including the extension of the model to the multiple question case. We develop a model for responses of the panel on each stage. Many questions are treated simultaneously and an autoregressive model is developed for explaining the responses of the group members as a function of the feedback. The errors are assumed to follow a matrix intraclass covariance structure. Marginal and conditional posterior distributions of the regression coefficient vector are found in both small and large samples. The broadly defined generic family of multidimensional Student-*t* distributions is found to play a major role in the results.

Keywords: BAYESIAN, MULTIVARIATE, GROUP DECISION MAKING, QUALITATIVE CONTROLLED FEEDBACK.

1. INTRODUCTION

Group judgment formulation and decision making using qualitative controlled feedback (QCF) was introduced in Press (1978). The work was extended to the multivariate case of many questions in Press (1980). In this paper we carry the work further by adopting the Bayesian point of view and developing the posterior distribution of the coefficient vector that relates individual responses of group members to explanatory variables.

The methodology was originally conceived in order to study how the U.S. Air Force might be reorganized. We will motivate the procedure, however, in a different context.

Suppose, for examples, a city planning bureau would like to resolve some public policy issues that are of importance to the city in various ways. They would like to determine how to allocate the resources in their budget so that "appropriate" funding is devoted to police, fire, and other municipal services, consistent with environmental considerations, political considerations, economic feasibility, engineering and scientific constraints, and perhaps other factors as well. These factors affect most people in some, possibly indirect, way, and no one person is likely to be knowledgable in all related areas.

It is decided to adopt a QCF procedure to assist the policy makers in generating the factors that argue for one allocation over another. A sample of panelists is taken from the city population; the panel members are each given a survey instrument that includes a battery of questions.

The survey instrument could be administered by mail, by telephone, by on-line computer, or whatever. The data collection protocol of QCF requires that each panelist respond to the questions independently of all other panelists, and without any panelists knowing the identity of any other panelists. Thus, the social pressures of face-to-face confrontation in a room, perhaps at the expense of logical reasoning, are avoided.

In applying a QCF procedure, each respondent is typically asked to answer a set of basic questions. In addition, the subject is asked to provide distinct reasons for each answer that will help justify the subject's answers. He will usually also be asked to answer some subsidiary questions that will serve to provide demographic and attitudinal information about the degree of expertise of the subject, his likely institutional biases, etc.

An intermediary is asked to collect all the answers. This person then forms a merged composite of the reasons provided by the panel for the answer to each question asked. This merging can be carried out with the aid of a computer editor. That is, in some situations this step may be carried out mechanically (if most reasons are listed in advance, panelists can check them off and a computer can talley them). Reasons can be coded and classified into some intrinsically orthogonal set (many reasons are probably just paraphrases of one another). The end product generated is a composite of reasons corresponding to each pair of questions and answers.

The composites of reasons are now presented to each panelist in a simple form (such as a checklist). Each panelist is then asked to answer the same set of questions a second time, only now, the panelist is exposed to the reasoning used by all other panelists. The numerical responses given by the other panelists are not provided for any subjects, nor do they receive any other data, such as sample group mean vectors. The composites of reasons are the only data fed back. As a result, the second stage response of a panelist is likely to differ from his first stage response only because he feels he has ignored some arguments used by other panelists. Note that panelists are not told the proportion of panelists who gave a particular reason; a panelist does not have any basis for deciding how much to weight each reason, in his own thinking, other than by adopting his own weighting system according to his own perceptions of value and importance.

This procedure is repeated until the process stabilizes, in the sense that respondents are not changing their responses very much from stage to stage.

There is room, however, for manipulation of the outcome by a devious intermediary who might misrepresent the composite fed back to the panel on each stage. This effect can be minimized by using a group of intermediaries to accomplish the task of forming a composite of reasons.

Earlier research involving group decision making and judgment formulation, and the effects of social interaction pressures, is summarized in Press, 1978. In Section 2 we develop a model for studying the relationships between responses to the questions, and the rationale the panel feels is most important to explain the answers. The model can also be used for predicting the next round's responses (in many situations, for economic or other reasons, it may be difficult or undesirable to carry out the process for one more stage).

The multiple question model is treated in greater detail from a sampling theory viewpoint in Press (1980). The methodology was applied to study a real problem in Press (1979b). Section 3 presents several distinct developments that provide methods for making Bayesian inferences useful for predicting the next round's responses. Finally, Section 4 provides a summary and conclusions.

2. MULTIPLE QUESTION MODEL

2.1 *First Stage*

Let $z_{in}(j)$ denote the numerical response of subject i, on stage n, to question j; $i = 1,2,\ldots,N$; $j = 1,2,\ldots,q$. Let $\mathbf{F}_n$ denote the totality of information obtained on stage n and feed back to each panelist at the beginning of stage $(n+1)$. Let $F^{(n)}$ denote the n-vector (F_j). Finally, let $\mathbf{X}{:}N\text{x}r$ denote a regressor matrix of explanatory variables observed for the N panelists (these are answers to subsidiary questions). Take $F^{(0)} \equiv \mathbf{0}$.

For the first stage model we adopt a simple regression with uncorrelated errors (subjects respond independently on the first stage). Accordingly, assume

$$\mathbf{z}_1(j)\,|\,\mathbf{X} = X\beta(j) + \mathbf{u}_1(j),$$
$$E(\mathbf{u}_1) = \mathbf{0},\ \text{var}[\mathbf{u}_1(j)] = \sigma_1^2(j)\mathbf{I}_N$$

where:

$$\mathbf{z}_1(j) = |z_{11}(j),\ldots,z_{N1}(j)]',\ \mathbf{u}_1 = [u_{11}(j),\ldots,u_{N1}(j)]'$$

$u_{i1}(j)$ denotes an error term, and $\beta(j)$ denotes an rx1 vector of unknown coefficients. For convenience, take

$$\underset{(\mathrm{Nxq})}{\mathbf{V}} = [\underset{(\mathrm{Nx1})}{\mathbf{u}_1(1)}, \ldots, \underset{(\mathrm{Nx1})}{\mathbf{u}_1(q)}], \qquad \underset{(\mathrm{qxN})}{\mathbf{V}'} = [\underset{(\mathrm{qx1})}{\mathbf{v}_1}, \ldots, \underset{(\mathrm{qx1})}{\mathbf{v}_N}],$$

and assume

$$E(\mathbf{v}_i\mathbf{v}_j) = \begin{cases} \Phi^*_i & i=j \\ \mathbf{0}, & i \neq j \end{cases}$$

If

$$\underset{(Nxq)}{Z_1} \equiv [\underset{(Nx1)}{z_1(1)}, \ldots, \underset{(Nx1)}{z_1(q)}], \qquad \underset{(rxq)}{\mathbf{B}} \equiv [\underset{(rx1)}{\beta(1)}, \ldots, \underset{(rx1)}{\beta(q)}]$$

the model may be written in the compact form

$$\underset{(Nxq)}{\mathbf{Z}_1} = \underset{(Nxr)}{\mathbf{X}} \; \underset{(rxq)}{\mathbf{B}} + \underset{(Nxq)}{\mathbf{V}} \; ; \tag{1}$$

where:

$$E(\mathbf{V}) = \mathbf{0}, \; \mathrm{cov}(\mathbf{v}_i, \mathbf{v}_j) = \mathbf{0}, \; i \neq j, \; \mathrm{var}(\mathbf{v}_i) = \Phi^*$$

The model of course represents a classical multivariate regression. The Gauss-Markov estimator of **B** is therefore

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_1 \tag{2}$$

2.2 *Feedback Stages ($n \geq 2$)*

For later stages, beyond the first, the model must change. This is because the composites of reasons fed back to each respondent cause their responses to be mutually correlated. Since they all get the same feedback, however, their responses on the next stage are likely to be *similarly* correlated (homogeneous, or intraclass correlation structure). Moreover, their answers on stage two are likely to be related to their answers on stage one. Adopt the autoregressive model

$$\triangle z_{in}(j) \equiv [z_{in}(j) \,|\, \mathbf{F}^{(n-1)}] - [z_{i,n-1}(j) \,|\, \mathbf{F}^{(n-2)}]$$

$$= \sum_{\alpha=1}^{\mathrm{R}_{n-1}(j)} c_i^{(\alpha)}(j)\,[1-\delta_{i,n-1}^{(\alpha)}(j)]\; p_n^{(\alpha)}(j) + u_{in}(j), \tag{3}$$

where $R_n(j)$ denotes the number of distinct reasons given by the panel (this is the number of reasons in the composite) for the answer to question j, on stage n; $\delta_{i,n}^{(\alpha)}$ is unity or zero, depending upon whether or not respondent i records reason α for his answer to question j, on stage n; $c_i^{(\alpha)}(j)$ is an unknown constant of proportionality (to be estimated); and $p_n^{(\alpha)}(j)$ denotes the proportion of respondents who record reason α for question j, on stage n (this will be interpreted as the weight or importance the panel gives to this reason). Note that even though the panel members do not know $p_n^{(\alpha)}(j)$, it can nevertheless be used in our model since the intermediary knows it or can compute it.

The model in eqn. (3) may be interpreted as follows; $\triangle z_{in}(j)$ represents the change in response for subject i, on question j, from stage (n-1) to stage (n). This change results from an incremental effect attributable to each reason (linear combination of effects). If the subject gave that reason on the last stage, there is of course no effect, while if he didn't give it, the effect is proportional to the importance of the reason (as measured by the proportion of panelists who gave the reason).

2.3 *Error Structure* ($n \geq 2$)

Define

$$\underset{(q\mathrm{x}1)}{\mathbf{u}_{in}} = [u_{in}(1), \ldots, u_{in}(q)]'$$

and assume

$$(1)\; E(\mathbf{u}_{in}) = \mathbf{0}\,,$$

$$(2)\; \mathrm{var}(\mathbf{u}_{in}) = \underset{(qxq)}{\Sigma_n} \; ;$$

$$(3)\; \mathrm{cov}(\mathbf{u}_{in}, \mathbf{u}_{jm}) = \begin{cases} \underset{(qxq)}{\Lambda_n}, \; i \neq j, & n = m, \\ \mathbf{0}, \; n \neq m \end{cases}$$

For compactness, let

$$\underset{(Nq\mathrm{x}1)}{\mathbf{u}_n} \equiv [\mathbf{u}'_{1n}, \ldots, \mathbf{u}'_{Nn}]' \; ;$$

Then, $E(\mathbf{u}_n) = \mathbf{0}$, and

$$\mathrm{var}(\mathbf{u}_n) = \underset{(Nq\mathrm{x}Nq)}{\Omega_n} = \begin{pmatrix} \Sigma_n & & \Lambda_n \\ & \ddots & \\ \Lambda_n & & \Sigma_n \end{pmatrix}$$

Ω_n is seen to be a matrix intraclass covariance matrix. Some of its properties are given, e.g., in Press, 1972, pp. 21, 48, 49 and in Press, 1979a. The assumption of equal diagonal blocks in Ω_n means we are assuming multivariate homoscedasticity. All off-diagonal elements of the qxq blocks are assumed to be identical (Λ_n). We are therefore assuming that in many situations it is reasonable to expect that the panel will be constituted with members who are sufficiently homogeneous in background so that a pattern of homogeneous correlation is reasonable.

2.4 *Transformations to Canonical Form*

Let

$$\underset{(q\mathrm{x}1)}{\Delta\mathbf{z}_{in}} = [\Delta z_{in}(1),\ldots,\Delta z_{in}(q)]' \quad ;$$

and assume

$$c_i^{(\alpha)}(j) = \underset{(1\mathrm{x}r)}{\mathbf{x}_i'} \; \underset{(r\mathrm{x}1)}{\mathbf{a}_\alpha(j)} \quad ; \qquad (4)$$

where $\mathbf{x}_i$ denotes the (rx1) vector of explanatory variables for subject i, and $\mathbf{a}_\alpha(j)$ denotes an (rx1) vector of unkown weights. For compactness, let

$$\underset{[R_{n-1}(j)\mathrm{x}1]}{\mathbf{c}_i(j)} = [c_i^{(1)}(j),\ldots,c_i^{(R_{n-1}(j))}(j)]' \quad ;$$

and

$$\underset{[rR_{n-1}(j)\mathrm{x}1]}{\mathbf{a}^{(n-1)}(j)} = [a_i'(j),\ldots,a'_{R_{n-1}(j)}(j)]' \quad ;$$

so that

$$\mathbf{c}_i(j) = (I \oplus x_i')\, a^{(n-1)}(j) \quad ,$$

where $\oplus$ denotes the direct product. We next combine all the observable explanatory data into one matrix. Define

$$\underset{[rR_{n-1}(j)\mathrm{x}1]}{\mathbf{w}_{in}(j)} = (\mathbf{I} \oplus \mathbf{x}_i)\,\mathbf{t}_{in}(j)$$

where:

$$t_{in}^{(\alpha)}(j) \equiv [1-\delta_{in-1}^{(\alpha)}(j)]\, p_n^{(\alpha)}(j) \quad ;$$

$$\underset{[1\mathrm{x}R_{n-1}(j)]}{\mathbf{t}_{in}'(j)} \equiv [t_{in}^{(1)}(j),\ldots,t_{in}^{(R-1(j))}(j)] \quad ;$$

and define

$$\underset{(q\mathrm{h}_{n-1})}{\mathbf{W}_{in}} = \begin{pmatrix} w_{in}'(1) & & 0 \\ & \ddots & \\ 0 & & w_{in}'(q) \end{pmatrix} \quad ;$$

and

$$\underset{(h_{n-1}\mathrm{x}1)}{\mathbf{a}^{(n-1)}} = [a^{(n-1)\prime}(1),\ldots,a^{(n-1)\prime}(q)]' \quad ;$$

where:

$$h_{n-1} \equiv r \sum^{q} R_{n-1}(j), \quad n \geq 2 \quad .$$

The model now becomes

$$\underset{(q\mathrm{x}1)}{\Delta\mathbf{z}_{in}} = \underset{(q\mathrm{x}h_{n-1})}{\mathbf{W}_{in}} \times \underset{(h_{n-1}\mathrm{x}1)}{\mathbf{a}^{(n-1)}} + \underset{(q\mathrm{x}1)}{\mathbf{u}_{in}} \qquad (5)$$

Combining all subjects, (5) becomes

$$\underset{(Nq\mathrm{x}1)}{\Delta\mathbf{z}_n} = \underset{(Nq\mathrm{x}h_{n-1})}{\mathbf{W}_n} \times \underset{(h_{n-1}\mathrm{x}1)}{\mathbf{a}^{(n-1)}} + \underset{(Nq\mathrm{x}1)}{\mathbf{u}_n} \quad ; \qquad (6)$$

where:

$$\Delta\mathbf{z}_n = (\Delta\mathbf{z}_{1n}',\ldots,\Delta\mathbf{z}_{Nn}')' \quad \mathbf{W}_n = (W_{1n}',\ldots,\mathbf{W}_{Nn}')' \quad ;$$

Iterating over the n stages gives

$$\underset{(Nq\mathrm{x}1)}{\mathbf{z}} \equiv \mathbf{z}_n\text{-}\mathbf{z}_1 = \underset{(Nq\mathrm{x}h)}{\mathbf{W}} \; \underset{(h\mathrm{x}1)}{\mathbf{a}} + \underset{(Nq\mathrm{x}1)}{\mathbf{u}} \quad ; \qquad (7)$$

where for $h = \sum_{j=1}^{n-1} h_j \quad n \geq 2$;

$$\underset{(h\text{x}1)}{\mathbf{a}} \equiv (\mathbf{a}^{(1)\prime}, \ldots, \mathbf{a}^{(n-1)\prime})' \quad ;$$

and

$$\underset{(Nq\text{x}h)}{\mathbf{W}} = (\mathbf{W}_2, \ldots, \mathbf{W}_n), \quad \underset{(Nq\text{x}1)}{\mathbf{u}} \equiv \sum_{j=1}^{n} \mathbf{u}_j$$

The transformed error vector in (7) satisfies

$$E(\mathbf{u}) = \mathbf{0}, \text{var}(\mathbf{u}) = \Omega = \begin{pmatrix} \Sigma & & \Lambda \\ & \ddots & \\ \Lambda & & \Sigma \end{pmatrix} \quad ; \tag{8}$$

where

$$\Sigma \equiv \sum_{j=2}^{n} (\Sigma_j), \qquad \Lambda \equiv \sum_{j=2}^{n} (\Lambda_j) \quad .$$

3. BAYESIAN INFERENCE

In this section we examine the unknown coefficient vector in the model defined by (7) and (8), from the Bayesian point of view. Four different approaches will be taken. First we will examine the coefficient vector conditional on the error covariance matrix. Then, we will develop an approximate conditional Bayesian estimator which is useful when samples are large. This approach ignores the intraclass structure of the covariance matrix and is useful for cases where the intraclass structure cannot be assumed. Next, in subsection 3, we will use the intraclass covariance structure when we develop the marginal posterior distribution of the coefficients. The result is complicated, and so a large sample solution is found. In the final subsection we develop a result which is useful in small samples.

3.1 *Known Covariance matrix*

From (7) and (8) it follows that under the assumption of normality on u, the density of the response vector (likelihood function) given the parameters and explanatory variables, is

$$p(\mathbf{z} \mid \mathbf{W}, \mathbf{a}, \Omega) \propto |\Omega|^{-1/2} \exp\{(-1/2)[(\mathbf{z}-\mathbf{W}\mathbf{a})'\Omega^{-1}(\mathbf{z}-\mathbf{W}\mathbf{a})]\} \quad . \tag{9}$$

Hence, if we adopt a vague prior for **a** (assuming Ω is known), its density is given by

$p(\mathbf{a}) \propto$ constant,

so that the posterior density is given (from Bayes theorem) by

$$p(\mathbf{a} \mid \mathbf{z}, \mathbf{W}, \Omega) \propto \exp\{(-1/2)[(\mathbf{z}-\mathbf{W}\mathbf{a})'\Omega^{-1}(\mathbf{z}-\mathbf{W}\mathbf{a})]\} \quad . \tag{10}$$

Note that we are using the common Bayesian convention of using the symbol $p(\cdot)$ to denote a generic density; the densities differ from one another according to the arguments and conditioning variables used.

Define the generalized least squares (and maximum likelihood) estimator

$$\bar{\mathbf{a}}(\Omega) = (\mathbf{W}'\Omega^{-1}\mathbf{W})^{-1}\mathbf{W}'\Omega^{-1}\mathbf{z} \quad . \tag{11}$$

Completing the square in the exponent in (10) shows that

$$p(\mathbf{a} \mid \mathbf{z}, \mathbf{W}, \Omega) \propto \exp\{(-1/2)[(\mathbf{a}-\bar{\mathbf{a}}(\Omega))'(\mathbf{W}'\Omega^{-1}\mathbf{W})(\mathbf{a}-\bar{\mathbf{a}}(\Omega)]\} \quad ;$$

so that

$$(\mathbf{a} \mid \mathbf{z}, \mathbf{W}, \Omega) \sim N[\bar{\mathbf{a}}(\Omega), (\mathbf{W}'\Omega^{-1}\mathbf{W})^{-1}] \quad . \tag{12}$$

That is, conditional on Ω, a posteriori, and adopting a vague prior on **a**, **a** is normally distributed, centered at the MLE, with precision matrix $(\mathbf{W}'\Omega^{-1}\mathbf{W})$.

We remark in passing that $\bar{\mathbf{a}}(\Omega)$ is the same estimator found from a frequentist point of view in Press (1979a).

3.2 *Large Sample Estimator*

One approximate large sample Bayesian estimator of a may be found (when Ω is unknown) by using the result obtained conditional on Ω, and then replacing Ω by a consistent estimator. This approach follows the spirit used in the frequentist analysis.

Suppose $\tilde{\Omega}$ is a consistent estimator of Ω (for unknown Ω). Then, the approximate posterior distribution of **a** is

$$(\mathbf{a} \mid \mathbf{z}, \mathbf{W}, \Omega) = N[\mathbf{a}(\tilde{\Omega}), (\mathbf{W}'\tilde{\Omega}^{-1}\mathbf{W})^{-1}] \quad .$$

A consistent estimator, $\tilde{\Omega}$, is developed in Press (1980). Thus, in large samples,

$$\mathbf{a}(\widetilde{\Omega}) \cong \mathbf{a}(\Omega) \quad ,$$

and **a** is approximately normally distributed.

3.3 *Marginal Distribution of* a

In this subsection we find Bayesian estimators based upon the marginal posterior distribution of *a*. The likelihood in (9) is equivalent to

$$(\mathbf{z}\,|\,\mathbf{W},\mathbf{a},\Omega) \sim N(\mathbf{W}\mathbf{a},\Omega) \quad , \tag{13}$$

where

$$\Omega = \begin{pmatrix} \Sigma & & \Lambda \\ & \ddots & \\ \Lambda & & \Sigma \end{pmatrix}$$

The posterior density of **a** is found by first reducing (13) to canonical form; then adopting a prior for the canonical form parameters, and finally applying Bayes theorem.

Define the orthogonal matrix $\Gamma = \Gamma_0 \oplus \mathbf{I}_q$, where Γ_0 denotes an orthogonal matrix of order N whose first row has equal elements. Then it is straightforward to check (see Press, 1979b, Theorem 5) that if

$$\Omega_0 \equiv \Gamma\Omega\Gamma' \quad ;$$

Ω_0 is block diagonal of the form

$$\Omega_0 = \begin{pmatrix} \Sigma_1 & & \mathbf{0} \\ & \Sigma_2. & \\ \mathbf{0} & & \Sigma_2 \end{pmatrix} ;$$

$\Sigma_1 = \Sigma + (N\text{-}1)\Lambda$, $\Sigma_2 = \Sigma\text{-}\Lambda$. Accordingly, define $\mathbf{z}^* = \Gamma\mathbf{z}$, $\mathbf{w}^* = \Gamma\mathbf{W}$. Then,

$$(\mathbf{z}^*\,|\,\mathbf{w}^*,\mathbf{a},\Omega_O) \sim N(\mathbf{w}^*\mathbf{a},\Omega_0) \tag{14}$$

We now view eqn. (14) as the canonical form of the problem and adopt $(\Sigma_1,\Sigma_2,\mathbf{a})$ as the canonical parameter set. Equivalently, if

$$\underset{(Nq\text{x}1)}{\mathbf{z}^*} \equiv (\underset{(1\text{x}q)}{\mathbf{z}_1^{*\,\prime}},\ldots,\underset{(1\text{x}q)}{\mathbf{z}_N^{*\prime}})' \quad ; \qquad \underset{(Nq\text{x}h)}{\mathbf{w}^*} \equiv (\underset{(h\text{x}q)}{\mathbf{w}_1^{*\,\prime}},\ldots,\underset{(h\text{x}q)}{\mathbf{w}_N^{*\prime}})' \quad ,$$

the canonical problem is the following:

$$(\mathbf{z}_1^*|\,\mathbf{w}_1^*,\mathbf{a},\Sigma_1) \sim N(\mathbf{w}_1^*\,\mathbf{a},\ \Sigma_1) \quad ;$$

$(\mathbf{z}_1^*, \ldots, \mathbf{z}_N^*)$ are independent and

$$(\mathbf{z}_j^*|\,\mathbf{w}_j^*,\mathbf{a},\Sigma_2) \sim N(\mathbf{w}_j^*\mathbf{a},\Sigma_2) \quad ,$$

for all $j = 2, \ldots, N$. A fundamental difficulty at this point is that Σ_1 depends on the sample size N (since $\Sigma_1 = \Sigma + (N\text{-}1)\Lambda$). To circumvent this difficulty we will seek a Bayesian solution to our problem which ignores one data point, namely, $\mathbf{z}_1^*$, and then we will seek a large sample solution, so that the loss of the one data point will be irrelevant.

Accordingly, we consider the joint posterior density

$$p(\mathbf{a},\Sigma_2\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) \propto \frac{p'(\mathbf{a},\Sigma_2)}{|\Sigma|^{(N-1)/2}}\; e^{-(1/2)tr\Sigma^{-1}B} \quad ;$$

where $p'(\mathbf{a},\Sigma_2)$ denotes the joint prior density of **a** and Σ_2,

$$\mathbf{B} \equiv \Sigma_{j=2}^{N}\,(\mathbf{w}_i^*\mathbf{a} - \mathbf{z}_j^*)(\mathbf{w}_i^*\mathbf{a} - z_j^*)'$$

and

$$\hat{\mathbf{z}} \equiv (\mathbf{z}_2^{*\prime},\ldots,\mathbf{z}_N^{*\prime})',\ \hat{\mathbf{w}} \equiv (\mathbf{w}_2^{*\prime},\ldots,\mathbf{w}_N^{*\prime})' \quad :$$

It is interesting to note that the sample covariance among the $(\mathbf{z}_2^*, \ldots, \mathbf{z}_N^*)$ vectors follows a non-central Wishart distribution.

Adopt the prior density

$$p'(\mathbf{a},\Sigma_2) = p_1'(\mathbf{a})\,p_2'(\Sigma_2) \quad ;$$

where:

$$p_1'(\mathbf{a}) \propto \text{constant},$$

$$p_2'(\Sigma_2) \propto \frac{1}{|\Sigma_2|^{n_0/2}}\, e^{-1/2tr\Sigma_2^{-1}G},\ \Sigma_2 > 0 \quad .$$

That is, the prior density of **a** is vague, and the prior density of Σ_2 is inverted Wishart. Note that $(\mathbf{G},\mathbf{n}_0)$ are assumed to be known hyperparameters. The posterior density now becomes

$$p(\mathbf{a},\Sigma_2\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) \propto \frac{1}{|\Sigma_2|^{(N+n_0-1)/2}}\, e^{-(1/2)tr\Sigma_2^{-1}(B+G)} \quad .$$

The marginal posterior density of $\mathbf{a}$ is found by integrating the joint density of $(\mathbf{a},\Sigma_2)$ with respect to Σ_2. Because of the known form of the inverted Wishart density, we readily effect the required integrarion and find

$$p(\mathbf{a}\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) \propto \frac{1}{|\mathbf{G}+\Sigma_{j=2}^{N}(\mathbf{w}_j^*\mathbf{a}\text{-}\mathbf{z}_j^*)(\mathbf{w}_j^*\mathbf{a}\text{-}\mathbf{z}_j^*)'\,|^{\nu/2}}, \tag{15}$$

where $\upsilon = N+n_0\text{-}q\text{-}2$. The posterior density in eqn. (15) is in the matrix-**T** family, but is quite complicated analytically. It could always be evaluated numerically, of course, but we seek instead a large sample approximation. An alternative approach will be developed for obtaining simple Bayesian results in small samples.

Large Sample Approximation

Let $\Phi \equiv (\mathbf{w}_2^*\mathbf{a}\text{-}\mathbf{z}_2^*,\ldots,\mathbf{w}_N^*\mathbf{a}\text{-}\mathbf{z}_N^*)$. Then, eqn. (15) becomes

$$p(\mathbf{a}\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) \propto |\mathbf{G}+\Phi\Phi'\,|^{-\nu/2} \propto |\mathbf{I}+\Phi'\mathbf{G}^{-1}\Phi\,|^{-\nu/2} \quad ;$$

or

$$p(\mathbf{a}\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) \propto \exp\{(-\upsilon/2)\log|\mathbf{I}_{N\text{-}1}+\Phi'\mathbf{G}^{-1}\Phi\,|\}$$

Let $(\lambda_1,\ \ldots,\ \lambda_{N\text{-}1})$ denote the latent roots of $\Phi'\mathbf{G}^{-1}\ \Phi$, and let $\mathbf{D} = \text{diag}\,(\lambda_1,\ldots,\lambda_{N\text{-}1})$. Then

$$\begin{aligned} p(\mathbf{a}\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) &\propto \exp\{(-\upsilon/2)\log|\mathbf{I}_{N\text{-}1}+\mathbf{D}_\lambda|\} \\ &= \exp\{(-\upsilon/2)\log\Pi_1^{N\text{-}1}(1+\lambda_i)\} \\ &= \exp\{(-\upsilon/2)\ \Sigma_1^{N\text{-}1}\log(1+\lambda_i)\} \end{aligned}$$

It will be shown shortly that λ_i decreases with sample size, N. Thus, for N sufficiently large, $|\lambda_i|\ll 1$, so that $\log(1+\lambda_i) \cong \lambda_i$. Then,

$$\begin{aligned} p(\mathbf{a}\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) &\propto \{(-\upsilon/2)\,\Sigma_1^{N\text{-}1}\lambda_i\} \\ &= \exp\{(-\upsilon/2)\,\text{tr}(\Phi'\mathbf{G}^{-1}\Phi)\} \\ &= \exp\{(-\upsilon/2)\,\Sigma_{j=2}^{N}(\mathbf{w}_j^*\mathbf{a}\text{-}\mathbf{z}_j^*)'\mathbf{G}^{-1}(\mathbf{w}_j^*\mathbf{a}\text{-}\mathbf{z}_j^*)\} \end{aligned}$$

Each term in the exponent is a quadratic form in $\mathbf{a}$. Combining terms gives

$$\begin{aligned} p(\mathbf{a}\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) \propto \exp\{-(½)[&\mathbf{a}'(\Sigma_2^N\mathbf{w}_i^{*\prime}(\mathbf{G}/\upsilon)^{-1}\mathbf{w}_j^*)\mathbf{a} - 2(\Sigma_2^N\mathbf{z}_i^{*\prime}(\mathbf{G}/\upsilon)^{-1}\mathbf{w}_j^*)\,\mathbf{a} \\ &+ (\Sigma_2^N\mathbf{z}_j^{*\prime}(\mathbf{G}/\upsilon)^{-1}\mathbf{z}_j^*)]\} \end{aligned}$$

To simplify, complete the square in $\mathbf{a}$ to get

$$p(\mathbf{a}\,|\,\hat{\mathbf{z}},\hat{\mathbf{w}}) \propto \exp\{(-½)[(\mathbf{a}\text{-}\alpha)'F(\mathbf{a}\text{-}\alpha)]\} \quad ,$$

where:

$$\alpha \equiv \mathbf{F}^{-1}\mathbf{b},\ \ \mathbf{b} \equiv \Sigma_2^N\mathbf{w}_j^{*\prime}+(\mathbf{G}/\upsilon)^{-1}\mathbf{z}_j^*,\ \ \mathbf{F} \equiv \Sigma_2^N\mathbf{w}_j^{*\prime}\,(\mathbf{G}/\upsilon)^{-1}\mathbf{w}_j^*$$

That is, a posteriori, in large samples,

$$\mathbf{a} \sim N(\alpha,\mathbf{F}^{-1}) \tag{16}$$

The only unfinished item remaining in this large sample approximation is to show that the latent roots of $\Phi'\mathbf{F}^{-1}\Phi$ go to zero with increasing sample size. The matrix $\Phi'\mathbf{F}^{-1}\Phi \equiv (r_{ij})$ where

$$r_{ij} \equiv (\mathbf{w}_i^*\mathbf{a}\text{-}\mathbf{z}_i^*)'\mathbf{G}^{-1}(\mathbf{w}_i^*\mathbf{a}\text{-}\mathbf{z}_j^*)$$

But

$$\underset{(Nqxh)}{\mathbf{w}^*} \equiv (\mathbf{w}_1^{*\prime},\ldots,\mathbf{w}_N^{*\prime})' = (\underset{(NxN)}{\Gamma_0} \oplus \mathbf{I}_q)\mathbf{w} \quad ;$$

where $\mathbf{w}$ only changes in dimension with increasing N. But Γ_0 is an orthogonal matrix each of whose elements is of order $N^{-1/2}$. So r_{ij} is of order N^{-1}. So its latent roots must vanish as $N \to \infty$.

Remark (1):

We note that since $\Sigma_1 = \Sigma+(N\text{-}1)\Lambda$, as N gets large, Σ_1 becomes very large, so $\mathbf{z}_1^*$ is less and less informative as $N\to\infty$. As a result, ignoring this observation is no great loss in large samples.

Remark (2):

The large sample Bayesian result shows that the elements of the regression coefficient vector $\mathbf{a}$ are, for large N, jointly normally distributed, so that inferences about particular coefficients are readily made.

Remark (3):

The large sample Bayesian result just found is meaningful when the number of subjects on the panel is large; the number of feedback stages may still be small.

3.4 *Small Samples*

To obtain a Bayesian result useful in small or moderate samples we adopt a different point of view than that used in subsection 3.3. Our approach now is to first ignore the (possibly) intraclass covariance structure in the likelihood function, but to recapture the structure in the prior distribution.

We begin with eqn. (14),

$$(\mathbf{z}^* \mid \mathbf{w}^*,\mathbf{a},\Omega_0) \sim N(\mathbf{w}^*\mathbf{a},\Omega_0) \tag{14}$$

Thus, the posterior distribution of $(\mathbf{a},\Omega_0)$ is

$$p(\mathbf{a},\Omega_0 \mid \mathbf{z}^*,\mathbf{w}^*) \propto \frac{p'(\mathbf{a},\Omega_0)}{|\Omega_o|^{1/2}} \exp\{(-\tfrac{1}{2})\mathrm{tr}\Omega_0^{-1}\mathbf{H}\} \tag{17}$$

where

$$\mathbf{H} \equiv (\mathbf{z}^*-\mathbf{w}^*\mathbf{a})(\mathbf{z}^*-\mathbf{w}^*\mathbf{a})'$$

and $p'(\mathbf{a},\Omega)$ denotes the prior density. Note that we are ignoring the intraclass structure of Ω_0 at this point.

For the prior density, assume $p'(\mathbf{a},\Omega_0) = p_1'(\mathbf{a})p_2'(\Omega_0)$, and

$$p_1'(\mathbf{a}) \propto \text{constant},$$

$$p_2'(\Omega_0) \propto \frac{1}{|\Omega_0|^{m/2}} \exp\{(-\tfrac{1}{2})\mathrm{tr}\Omega_0^{-1}\mathbf{M}\}$$

where $(m,\mathbf{M})$ are assumed to be known hyperparameters, $\mathbf{M}>0$.

The joint posterior density becomes

$$p(\mathbf{a},\Omega_0 \mid \mathbf{z}^*,\mathbf{w}^*) \propto \frac{1}{|\Omega_0|^{(m+1)/2}} \exp\{(-\tfrac{1}{2})\mathrm{tr}\Omega_0^{-1}(\mathbf{M}+\mathbf{H}\}$$

In Ω_0 this expression is the kernel of an inverted Wishart distribution so it is readily integrated to give the marginal density

$$p(\mathbf{a} \mid \mathbf{z}^*,\mathbf{w}^*) \propto \frac{1}{|\mathbf{M}+\mathbf{H}|^{(m-Nq)/2}},$$

or

$$p(\mathbf{a} \mid \mathbf{z}^*,\mathbf{w}^*) \propto \frac{1}{\{1+(\mathbf{w}^*\mathbf{a}-\mathbf{z})'M^{-1}(\mathbf{w}^*\mathbf{a}-\mathbf{z}^*)\}^{(m-Nq)/2}}$$

Completing the square in **a** gives

$$p(\mathbf{a} \mid \mathbf{z}^*,\mathbf{w}^*) \propto \frac{1}{\{\upsilon^*+(\mathbf{a}-\alpha^*)'\mathbf{Q}^{-1}(\mathbf{a}-\alpha^*)\}^{(\upsilon^*+h)/2}}, \tag{18}$$

where:

$$\alpha^* \equiv (\mathbf{w}^{*\prime}\mathbf{M}^{-1}\mathbf{w}^*)^{-1}(\mathbf{w}^{*\prime}\mathbf{M}^{-1}\mathbf{z}^*),\ \upsilon^* \equiv m\text{-}Nq\text{-}h,$$

$$\mathbf{Q}^{-i} \equiv (\mathbf{w}^{*\prime}\mathbf{M}^{-1}\mathbf{w}^*)\upsilon^*/\beta^*$$

$$\beta^* \equiv 1+\mathbf{z}^{*\prime}\mathbf{M}^{-1}\mathbf{z}^*\text{-}\alpha^{*\prime}\mathbf{w}^{*\prime}\mathbf{M}^{-1}\mathbf{w}^*\alpha^*$$

That is, **a** follows an *h-dimensional Student* **t**-density with mean α^*, and υ^* degrees of freedom. Then,

$$\begin{aligned} E(\mathbf{a} \mid \mathbf{z}^*,\mathbf{w}^*) &= \alpha^* = (\mathbf{w}^{*\prime}\mathbf{M}^{-1}\mathbf{w}^*)^{-1}(\mathbf{w}^{*\prime}\mathbf{M}^{-1}\mathbf{z}^*) \\ \mathrm{var}(\mathbf{a} \mid \mathbf{z}^*,\mathbf{w}^*) &= (\upsilon^*/\upsilon^*\text{-}2)\mathbf{Q} \\ \mathrm{var}(\mathbf{a} \mid \mathbf{z}^*,\mathbf{w}^*) &= (\beta^*/(\upsilon^*\text{-}2))(\mathbf{w}^{*\prime}\mathbf{M}^{-1}\mathbf{w}^*)^{-i} \end{aligned} \tag{19}$$

Discussion of Prior

The mean and variance of the inverted Wishart distribution are well known (see e.g. Press, 1972, p. 111). Therefore

$$E(\Omega_0) = \mathbf{M}/(m\text{-}2Nq\text{-}2)$$

But if we subjectively believe that

$$\Omega_0 = \begin{pmatrix} \Sigma_1 & & 0 \\ & \Sigma_2 & \\ 0 & & \ddots\ \Sigma_2 \end{pmatrix},$$

we should take

$$\mathbf{M} \equiv \begin{pmatrix} \mathbf{M}_1 & & 0 \\ & \mathbf{M}_2 & \\ 0 & & \ddots\ \mathbf{M}_2 \end{pmatrix}, \mathbf{M}_1 > 0, \mathbf{M}_2 > 0.$$

Then,

$$E(\Sigma_1) \equiv \frac{\mathbf{M}_1}{m\text{-}2Nq\text{-}2}, \qquad E(\Sigma_2) = \frac{\mathbf{M}_2}{m\text{-}2Nq\text{-}2},$$

and $E(\Omega_0) = \mathbf{0}$ for all elements of Ω_0 not in the block diagonal elements. Moreover, if $\Omega_0 \equiv (\omega_{\alpha\beta})$, for all (α,β) not in the block diagonal elements

$$\mathrm{var}(\omega_{\alpha\beta}) = \frac{m_{\alpha\alpha}m_{\beta\beta}}{(m\text{-}2Nq\text{-}1)(m\text{-}2Nq\text{-}2)(m\text{-}2Nq\text{-}4)} ;$$

where $\mathbf{M} \equiv (m_{\alpha\beta})$. Note that $\mathrm{var}(\omega_{\alpha\beta})$ is of order m^3; that is, $\mathrm{var}(\omega_{\alpha\beta})$ goes to zero with increasing m^3. We can always choose m large enough so that all elements off the block diagonal elements of Ω_0 are centered at zero, with very small variance. Note from eqn. (19) that $\mathrm{var}(\mathbf{a}\,|\,\mathbf{z}^*,\mathbf{w}^*)$ goes to zero with increasing υ^* (which is linear in m). By selecting $(\mathbf{M}_1,\mathbf{M}_2)$ appropriately, and choosing m sufficiently large this prior distribution will be sufficiently rich to accommodate many classes of subjective information.

This type of prior is not recommended for the general case, since the structure of the prior distribution is too restrictive[1]. Our reasoning is that although elements of Ω_0 not in the blocks on the main diagonal are centered at zero with arbitrarily small variance, because there are only two parameters in the inverted Wishart distribution, viz. $(\mathbf{M},m)$, the elements of Ω_0 that do lie in the main diagonal blocks are simultaneously constrained in all of their moments (by taking m large). Such constraints may not always be desirable. For the general case, an alternative prior for Ω_0 which is richer in parameters is recommended. We propose such a prior below.

Generalized Prior Distribution

A generalized family of Wishart type distributions was introduced by Roux, 1971. The generalization includes hypergeometric functions of matrix argument. A form of the associated density which widens the parameter spaces is given (for a general *pds* matrix $\hat{\mathbf{X}}$) by

$$f(\hat{\mathbf{X}}) = c\,|\hat{\mathbf{X}}|^{\gamma-(q+1)/2} \exp\{-\mathrm{tr}(\mathbf{J}\hat{\mathbf{X}})\}\, {}_{r}{}^{*}F_{q}{}^{*}(\delta;\eta;\mathbf{JRJ}\hat{\mathbf{X}}) \quad ; \qquad (20)$$

for $\hat{\mathbf{X}}$: qxq, $\hat{\mathbf{X}}>0$, $\delta \equiv (\delta_1,\ldots,\delta_{r}{}^{*})'$, $\eta \equiv (\eta_1,\ldots,\eta_{q}{}^{*})'$, $\mathbf{J}$: qxq, $\mathbf{R}$: qxq, $\mathbf{J}>0$, $\mathbf{R}>0$, and ${}_{r}{}^{*}F_{q}{}^{*}(\bullet)$ denotes the generalized hypergeometric function of matrix argument (see Constantine, 1963). The normalizing constant is given by

$$c = \frac{|\mathbf{J}|^{\gamma}}{\Gamma_q(\gamma)\,{}_{(r^*+1)}F_{q}{}^{*}(\delta,\gamma;\eta;\mathbf{JR})} ,$$

[1] The restrictiveness of the structural form of the Inverted Wishart distribution has already been noted by Rothenberg, 1963, in a different context (see References).

where $\Gamma_q(\gamma)$ denotes a q-dimensional gamma function. The parameters (δ_i,η_j), $i=1,\ldots,r^*$; $j=1,\ldots,q^*$, are restricted to take those values for which $f(\hat{\mathbf{X}})$ is positive.

Now let $\Omega_0 \equiv \hat{\mathbf{X}}^{-1}$, replace q by Nq (the dimension of Ω_0), and transform the density in (20) to yield the generalized inverted Wishart density

$$p_2'(\Omega_0) = \frac{c}{|\Omega_0|^{\gamma+(Nq+1)/2}} \exp\{-\mathrm{tr}(\mathbf{J}\Omega_0^{-1})\}\, {}_{r}{}^{*}F_{q}{}^{*}(\delta;\eta;\mathbf{JRJ}\Omega_0^{-1}) \qquad (21)$$

Using eqn. (21) in eqn. (17), with $p'(\mathbf{a},\Omega_0) \propto p_2'(\Omega_0)$, gives the joint posterior density

$$p(\mathbf{a},\Omega_0\,|\,\mathbf{z}^*,\mathbf{w}^*) \propto |\Omega_0|^{-(2\gamma+Nq+2)/2} \exp\{(-\tfrac{1}{2})\mathrm{tr}\Omega_0^{-1}(\mathbf{H}+2\mathbf{J})\}$$

$$\bullet\ {}_{r}{}^{*}F_{q}{}^{*}(\delta;\mathbf{h};\mathbf{JRJ}\Omega_0^{-1}). \qquad (22)$$

The marginal posterior density of $\mathbf{a}$ is found by integrating (22) with respect to Ω_0. The integration is carried out by reference to eqn. (21), using its normalizing constant. The result is

$$p(\mathbf{a}\,|\,\mathbf{z}^*,\mathbf{w}^*) \propto \frac{1}{|2\mathbf{J}+(\mathbf{z}^*\text{-}\mathbf{w}^*\mathbf{a})(\mathbf{z}^*\text{-}\mathbf{w}^*\mathbf{a})'|^{\gamma+1/2}}$$

$$\bullet\ {}_{(r^*+1)}F_{q}{}^{*}(\delta,\gamma+\tfrac{1}{2};\eta;\mathbf{JRJ}(2\mathbf{J}+\mathbf{H})^{-1}) \quad ; \qquad (23)$$

where: $\mathbf{H} \equiv (\mathbf{z}^*\text{-}\mathbf{w}^*\mathbf{a})(\mathbf{z}^*\text{-}\mathbf{w}^*\mathbf{a})'$. If we identify $\mathbf{M} \equiv 2\mathbf{J}$, $2\gamma \equiv m\text{-}Nq\text{-}1$, and take $\mathbf{R} \equiv \mathbf{0}$, it is immediately seen that the result obtained in (18), for the inverted Wishart prior, is a special case of eqn. (23). This result, however, has the advantage of being richer in parameters and can therefore accommodate a much greater variety of types of subjective information. Inferences about $\mathbf{a}$, however, are more complicated, and will require the use of zonal polynomial tables in order to evaluate the hypergeometric functions in (23) (see James and Parkhurst, 1974). The parameters of the hypergeometric functions are selected so as to satisfy the block diagonal structure of Ω_0.

4. CONCLUSIONS AND SUMMARY

The qualitative controlled feedback process of forming group judgments and making decisions has been examined from a Bayesian viewpoint. The group responses to many questions was modeled as an autoregressive process with coefficient vector $\mathbf{a}$.

It was shown that if the error covariance matrix, Ω, is known, the posterior distribution of **a** is normal, and centered at the generalized LSE. In large samples, if Ω is unknown, a consistent estimator may be used to make conditional inferences about **a**.

Bayesian inferences can also be made marginally, without reference to Ω. Assuming intraclass covariance structure, the marginal posterior distribution of **a** was shown to be, approximately, a complicated member of the matrix T family of distributions. We developed a normal distribution approximation which is very useful in large samples, however. For small sample situations involving the intraclass covariance structured situation we developed a posterior multivariate Student t-density for **a**. This result although useful for many situations is somewhat restrictive in the types of prior information it will accommodate. A more general result was obtained using generalized inverted Wishart distribution priors. The result is more complicated to use, however.

Finally, note that the entire QCF process is subjective in nature. It is therefore not surprising that inferences about the relationship between the responses of the panel members, and their individual characteristics and judgmental behavior regarding the reasons other panelists give, would depend heavily upon the nature and quantity of the prior information available.

ACKNOWLEDGMENT

I would like to thank Dr. Hiroshi Watanabe for his careful reading of the manuscript. I am also grateful to Professors J.R. Dunsmore and S. Geisser for helpful comments.

REFERENCES

CONSTANTINE, A.G. (1963). Some Non-central Distribution Problems in Multivariate Analysis, *Ann. Math. Statist.*, **34** 1270-1285.

JAMES, A.T. and A.M. PARKHURST (1974). *Selected Tables in Mathematical Statistics,* Vol. **2**. (H.L. Harter and D.B. Owen, eds.) California: Institute of Mathematical Statistics.

PRESS, S. JAMES (1972). *Applied Multivariate Analysis.* New York: Holt, Rinehart and Winston

— (1978). Qualitative Controlled Feedback for Forming Group Judgments and Making Decisions. *J. Amer. Statist. Assoc.* **73**, 526-535.

— (1980). Multivariate Group Judgments by Qualitative Controlled Feedback, In *Multivariate Analysis V,* (P.R. Krishnaiah, ed.) New York: North Holland, 581-591.

— (1979a). Matrix Intraclass Covariance matrices with Applications in Agriculture, *Tech. Rep.* **49**. Department of Statistics, University of California, Riverside.

— M.W. ALI, and E. YANG (1979b). An empirical Study of a New Method for Forming Group Judgments: Qualitative Controlled Feedback, *Technological Forecasting and Social Change,* **15**, 171-189.

ROTHENBERG, T.J. (1963) A Bayesian Analysis of Simultaneous Equation Systems, *Tech. Rep.* **6315**, Econometric Institute, Rotterdam.

ROUX, J.J.J. (1971) On Generalized Multivariate Distributions, *South African Statist,* **5**, 91-100.

DISCUSSION

I.R. DUNSMORE *(University of Sheffield)*:

We heard in the discussions yesterday a prediction that Bayesian statistics would be dead by the end of the twentieth century. Many of us may still be around then to ensure that this will not be the case; and some of us are going about this task by concentrating attention on predictive distributions of future observations rather than on posterior distributions of parameters. A welcome step in the right direction therefore is this interesting and clearly presented paper by Professor DeGroot in which he attempts to model how a statistician proposes predictive distributions in a sequence of similar decision problems. Can we apply the method practically?

After hearing and assessing all the information at my disposal my initial (predictive) statement is that the theory is beautifully modelled but from the practical viewpoint I am dubious of its worth. Yet Professor DeGroot is a leading authority, and so following his doctrine of probabilism or probabiliorism that a doctrine recognized by a leading authority to be correct can be taken to be correct, the reality must be that I am subject to appreciable specification bias and that what has been presented lies well out in the tails of my initial predictive distribution. So I must learn to do better and reexamine the information available.

My main stumbling point is the separation of "previous information" from "tendency to mis-specify". Consider the problem S_1 in the location-shift model. After duly considering *all* his available information the statistician specifies a predictive distribution for X_1, say $N(M_1.T_1)$. Professor DeGroot argues that we could arrive at this distribution by modelling the statistician's behaviour as follows. He is just about to specify that X_1 is $N(m_1,r_1)$ when he remembers his specification bias θ_1, and so takes X_1 as $N(m_1+\theta_1/r_1^{1/2},r_1)$. This of course presumes that he can separate his information for assessing $N(m_1,r_1)$ from his specification bias information. With a prior $N(\mu_1,\tau_1)$ on θ_1 this distribution averages out to a predictive assessment of $N(M_1 = m_1+\mu_1/r_1^{1/2},T_1 = (\tau_1/(1+\tau_1))r_1)$. Now move on to problem S_2. The statistician will learn something from observing the outcome of S_1 about the type of specification errors he is likely to make and, consciously or subconsciously, he will use this knowledge about his own personality when he initially specifies the predictive distribution of X_2 as $N(M_2,T_2)$. Professor DeGroot then suggests the same procedure for modelling the learning by arguing that the statistician is just about to specify a $N(m_2,r_2)$ distribution for X_2 when he remembers his second specification bias θ_2, with prior $N(\mu_2,\tau_2)$, so that he gives his predictive distribution as $N(M_2 = m_2+\mu_2/r_2^{1/2},T_2 = (\tau_2/(\tau_2+1))r_2)$. But here θ_2 must measure in the learning process that statistician's "failure to correct sufficiently" in specifying m_2 and r_2. These seem to me to be much more nebulous quantities to deal with, and it is most doubtful if the assessment of m_2 and r_2 is now independent of the information for assessing θ_2.

For the location shift parameter Professor DeGroot models his procedure most elegantly, but I am much less convinced with the modelling of the scale or precision example of §6. There does not seem to me to be any intuitive appeal in the learning process $\alpha_2 = \alpha_1'+\delta, \beta_2 = \beta_1'+\delta$ when R_2 somehow measures the statistician's "failure to correct sufficiently". The problem presumably intensifies for the location and precision shift problem.

One final comment concerns the element of distribution dependence in the arguments presented. No observation will allow the statistician to move away from a normal distribution (or a Student distribution) for example. Might it not be that if 80% of your observations lie more than 5 standard deviations below the mean of your predicted distribution then skewed predictive distributions may be more appropriate.

Turning now to the second paper, my first task is to thank Professor Press for his interesting talk today. The Bayesian modelling of his qualitative controlled feedback ideas was clearly the next progression in his study of group judgment formulation - althought many here might presumably have expected it to be the first step. Professor Press has presented us with several different Bayesian approximate models via some elegant and intricate matrix manipulation.

My second task is not as easy. The question I wish to pose is "What does it all mean from a practical point of view?". The practical relevance and interpretation of the posterior distribution for **a** somewhat defeats me. For example, just consider the dimensionality of **a**. This is an $hx1$ vector in the model (7) where $h = \sum_{\ell=1}^{n-1} h_\ell = \sum_{\ell=1}^{n-1} r \sum_{j=1}^{q} R_\ell(j)$. So if, for example, there are $r=5$ explanatory variables, $q=2$ questions, and $R_1(1)=4$, $R_1(2)=5$, $R_2(1)=6$, $R_2(2)=9$, then by the third stage h is already 120. Another point is that some of the reasons within the list $R_2(1)$, for example, will be contained in the list $R_1(1)$. Is it then necessary to have different $a_\alpha(1)$'s for stages 1 and 2 or could we allow the $p_n^{(\alpha)}(j)$'s in the model to account for the variability over n? (Although the notation is not explicit on this fact it seems that the a's are considered to vary with n).

It is also clear that the adequacy or otherwise of this ingenious model must be thoroughly investigated. One question on this count that I would like to pose is to ask if the model can cope with someone who "about turns" in his answer or opinion even with the same reasons, which after all is a common tactic of some committee members!

Turning briefly to the approximations to the posterior distributions I have one comment on the large-sample approximation of §3.3. Is it sensible to ignore the variable with the "largest" variance matrix? To an outsider it must look as though your conclusions will be more accurate than they should have been.

My third and final task is simply a plea. Please may we see this elegant statistical theory transformed into useful statistical practice.

S. GEISSER (*University of Minnesota*):

Professor DeGroot, no doubt manifesting his great flair for the sensibilities of our Spanish hosts, presented this paper because it implements what that renowned Spanish and American philosopher, Santayana, rather broadly intimated that those who ignore the data are doomed to repeat the mistakes of the past.

DeGroot has elegantly formulated a Bayesian apparatus that might serve to dampen and eventually avoid subjective biases. In some respects, he is more vitally concerned with the non-Bayesian aspects of the problem. He lists several possible

alternative conclusions to be drawn from the fact that a subjectively predicted value is far from where expected:

1) The subjective predictive distribution is misleading (model doesn't fit).
2) A rare event has occurred because of an unusual observation (all we have to do is stick the conjunction "or" between (1) and (2) and we have Fisher's so-called logical disjunction).

What is left then is the last alternative which states:

3) There is no relation between the predictive distribution and the observed value.

I would call (3) the archsubjectivistic view, but I refuse to pay it much heed because if anything could persuade me to turn in my Bayesian credentials, it is this extreme view. And, of course, DeGroot is also too sensible to accept this. (What then is the point of a predictive distribution if it is to bear no relation to an observation?) At any rate, he claims to model what he terms the "behaviour" of the statistician. (Wasn't it Neyman who coined the term inductive behaviour?) It would appear that the accretions of the past are not so easily disposed of on the ash heap of history, and perhaps rightly so. In modelling the situation he finds he must explore the mind-boggling hyperworld of hyperparameters. But DeGroot is a slippery Bayesian and he refuses to assume an extreme position by not trying to n-guess himself, thus extricating himself from the hyperparametric thicket he has created.

He sensibly assumes that his specification error should be smaller the second time by a fraction of the first expected specification error -and so it goes, recursively that is.

There is also the hyperworld of γ itself, the DeGroot rate of learning parameter which is also subjected to a distribution and as decisions made and observations obtained he learns about his learning rate —and he learns about how he learns about his learning rate— a veritable underworld of learning similar to Fisher's underworld of probability.

It seems that this method of "learning" could easily be called the "DeGroot Opinion Processor and Evaluator" whose descriptive acronym aptly describes the learner. It is a smooth "rational" method that has none of the qualities of human learning at its best - inspiration, acuity, perception, and concentration - and is better suited for plodding, dull, and unimaginative automatons.

In the light of all this perhaps mildly unfair criticism, let me also give Professor DeGroot something less amorphous to which he can respond. Consider the predictive subjective model which is normal and unbiased, but may be more widely dispersed than it should be, whatever that may mean, then how important is it for the statistician who guesses only the mean value of his subjective distribution, and if so, how concerned should he be about having too large a subjective misspecification variance if he only will guess a single value?

Also, couldn't the statisticians' misspecification bias really be due to a misreading of auxiliary conditions which may affect the payoff of investment decisions, and instead of smoothly adjusting his parameters, he may want to radically reconsider his whole set-up after a few "bad" decisions.

In conclusion, let me say that I thoroughly enjoyed this paper as it compelled me to consider how important it is for a Bayesian to become involved in the human learning process and what a giant step DeGroot has taken in grappling with this problem and developing a point of view which is certainly not entirely orthogonal to the truth, if there is any here.

Professor Press is to be congratulated on his usual virtuoso performance in manipulating distributions of random matrices. But the refrain "What's it all about Alfie?" keeps coming to mind. If our good friend the new socialist mayor of Valencia wanted to resolve a pressing public policy issue -say the building of a Bayesian conference Center- why should he use Qualitative Controlled Feedback rather than having an open discussion and a popular referendum?

Another difficulty that I have previously pointed out about feedback procedures is the potential for misuse by a devious intermediary who would feedback false or slanted information in order to manipulate the outcome.

J.M. BERNARDO (*Universidad de Valencia*):

The nice mathematical properties associated to the combination of a normal model with the inverted Wishart distribution used by Professor Press in his prior specification have been exploited in a number of Bayesian papers. However, as he points out, this may be too restrictive. I would like Professor Press to expand on this point, making explicit the type of situations for which he feels this prior might be sensible, and commenting whether he knows of any real life applications. Information about possible interactive computer routines for this type of prior specification would be valuable.

A.P. DAWID (*The City University*)

Professor DeGroot is surely right to argue that the Bayesian should be ready to confront his internal probabilistic view of the world with some external reality, and to modify his view, rather than the real world, if there appears to be a conflict. One of the weaknesses of subjectivist theory, confined as it is in its comfortable coherent cocoon, is that it does not seem to make any formal allowance for such a confrontation. Something can be said, however.

Suppose that a weather forecaster has to make, each day, a statement of his probability of precipitation within a specified 12 hour period of the next 24 hours (Murphy and Winkler, 1977). He need not have any model in mind, but is merely stating his conditional probability of "rain tomorrow", given his whole knowledge today. Let us now consider all those days for which his forecast probability lay in the range, say, $1/3 \pm \epsilon$, and suppose that the number of such days is (conceptually) infinite. Then, using martingale theory, one can show that the limiting relative frequency of rain on such days lies, with probability one, in the same range $1/3 \pm \epsilon$. The probability referred to here is, of course, that corresponding to the forecaster's subjective opinions.

Similarly, if each day he gives a *credible interval* which he assesses to have, say, 50% probability of containing tomorrow's maximum temperature, then he should believe, with probability one, that in the limit 50% of such intervals will contain the true value.

Note particularly that the above theory does not require any assumption of independent or "unrelated" problems, merely that each forecast be made in the light of full knowledge of the outcomes of previous forecasts. So, in a sense, the Bayesian is out-frequencing the frequentist.

Now suppose that, in a very long sequence of such forecasts, only 30% of the forecaster's 50% credible intervals are covering their true values; then an event has occurred to which the forecaster assigned very low probability. It seems to me clear that the world is telling the forecaster that his Bayesian beliefs, coherent though they may be, are out of touch with reality. However this logic is squarely in the spirit of significance testing (or of Professor Box's contribution to this conference) and I cannot see how to justify it from the position of the self-contained subjectivist.

The above considerations apply for the forecaster's "true" probabilities. It is easy for him to cheat, by quoting probabilities in which he does not really believe, so as to appear "well-calibrated" (DeGroot, 1979). Moreover, even if his true probabilities are well-calibrated, this does not necessarily mean that they are "accurate" in all respects; and even if they are accurate, they may not be of much *substantive* value if the forecaster is a poor meteorologist.

Professor DeGroot is working in the following framework. The forecaster sets up a mathematical model which, he hopes, is an adequate approximation to his true internal beliefs, which he in turn hopes correspond, somehow, to the real world. But the real world says "Not so". So the forecaster replaces his initial model with a more complex one, which he hopes will lead to more "accurate" forecasts. Clearly the process can be iterated, and now bears a very close resemblance to Box's cycle of estimation and criticism.

But I believe there is a danger of falling into an infinite regress. However much we refine our subjective models, or learn about our learning process, the real world may still surprise us by throwing up events which we believe shouldn't occur. So in what sense, if any, have we improved our probability modelling?

W.H. DUMOUCHEL (*Massachussets Institute of Technology*):

Professor DeGroot should be thanked for tackling the somewhat taboo question: "What can a Bayesian do who is consistently wrong?" I would like to suggest another possible approach and solution.

Suppose the statistician encounters a sequence of trials in which it is necessary to predict a continuous variable S, after which the observation $X = x_n$ is made at the n^{th} trial. Let F_n be the statistician's predictive distribution function for X just before the n^{th} trial, and let

$$U_n = F_n(x_n) \quad ,$$

which is observable after the n^{th} trial. Now, if the statistician computes F_n correctly just before the n^{th} trial, the sequence, U_1, U_2, ... should be indistinguishable from a sequence of independent uniform variables on (0,1). If the predictive distribution F_n is being consistently computed incorrectly, then it may be that $\{U_n\}$ behaves like a sequence of i.i.d. variables from some unknown distribution G. A possible assumption is that G is a beta distribution with parameters α and β. Then the question of whether the statistician is consistently wrong in computing F_n for the predictive distribution of X_n boils down to an hypothesis about α and β, where the variable $U_n = F_n(X_n)$ has a Beta (α,β) distribution, i.e,

$$H\colon \alpha = \beta = 1$$
$$\bar{H}\colon (\alpha,\beta) \neq (1,1)$$

This problem can be treated as a sharp null hypothesis problem. The statistician formulates a prior distribution on (α,β) with

Ω_0 = prior odds in favor of H
Ω_n = odds in favor of H conditional on $U_1, U_2, \ldots, U_n$

Then the predictive distribution of U_{n+1} is $\hat{G}_n(u) = \int G(u;\alpha,\beta)dP_n(\alpha,\beta)$, where $G(u;\alpha,\beta)$ is the beta distribution function, and P_n is the posterior distribution of (α,β).

My proposed decision rule is then:

1. Make no corrections to inferences about X as long as $\Omega_n \geq 1$(or $\Omega \geq k$).
2. If $\Omega_n < 1$, then correct F_{n+1} to make U_{n+1} uniform. That is, $U_{n+1} = F^*_{n+1}(x) = \hat{G}_n(F_{n+1}(X))$. The 100 u percentile of x_{n+1} is $F^{-1}_{n+1}\ \hat{G}^{-1}_n(u)$. In general, for $k \geq 1$, replace F_{n+k} by $\hat{G}_n F_{n+k}$ in all inferences about X.
3. Whenever $\Omega_n < 1$ and step 2 above has been taken, start over with a new reassessment of $\Omega_0, P_0(\alpha,\beta)$, etc. corresponding to the new definitions of F and U.

If the assumption that $\{U_n\}$ is approximately a sample from some beta distribution is correct, then when n is large, $\hat{G}_n F_{n+1}$ will produce just the right correction to F_{n+1}, as simple calculations show. Of course it is no simple calculation to compute Ω_n or $\hat{G}_n$, which depend on the choice of prior $P_0(\alpha,\beta)$ for α and β when n is small, but less so when n is large.

Although this method poses computational problems, it is very general, being applicable to any continuous predictive situation, and it provides a method for simultaneously correcting for error in scale and location, since the two parameter (α,β) are available for the estimation of G, and even more general families could be used instead of the beta family. I hope to develop this method in future work.

S. FRENCH (*University of Manchester*):

I should like to comment on Professor Press' paper. It seems particularly important to emphasise a point that was clear from Professor Press' presentation at the conference, but not clear from his written paper. At least, I for one was misled. The methods of this paper are directed at the problem of gathering and summarising group opinion for a decision maker *exterior* to the group. They should not be considered as methods to help a group of decision makers reach consensus amongst themselves. That no such methods can exist should now be well known, Arrow (1963), Luce and Raiffa (1958), Patternaik (1978), French (1980). I say "should be" since I am aware that some decision analysts see their task of advising a group as one of generating a group probability distribution and a group utility function and then of advising the action with maximum group expected utility. Such analyses are unlikely to be rational in the Bayesian sense. Professor Press' methods appear tailor-made for such "irrational" analyses. I was glad to hear from his presentation that such a close fit was unintentional.

Turning now to the correct use of Professor Press' analysis, I am far from convinced that anonymity will lead to "objetivity". I personally judge a person's opinions and his reasons for holding those opinions against the background of his character. Moreover, I am aware that some of the best opinions are held without the holder being able to express why he holds them. Consider a firm taking advice from a group of experts within a research and development department. How will Professor Press' method assimilate the opinions of a man with a hunch. By definition he cannot articulate his reasons for his opinion. So his view will not communicate itself to the rest of the group. Yet, the rest may all agree "Old Charlie has a gut feeling for winning projects. If he says it's a winner then that's good enough for me". This example is contrived maybe, but I hope it makes my point. One gathers information through a group of experts rather than the literature, when it is clear that there are too many uncertainties involved for them to be objectively analysed. Thus one intentionally asks the panel to use their intuitive expertise. Yet this method concentrates their attention on that part of their judgemental process which they can articulate, but not necessarily directly upon the part for which they were employed.

It may well be that Professor Press does not see this method as being used to sample expert opinion, but rather a large population of consumers. His paper does indeed concentrate on an example where a city planning bureau surveys public opinion. (However, see Harman and Press (1978)). Here too, I am worried about the applicability of his methods. A sample survey is meant to be representative of the population sampled. Yet it is a basic property of qualitative controlled feedback that it changes the initial opinions of the group. So the output of a sample survey conducted by Professor Press' methods is unlikely to be representative of the opinions sampled. At the end of the analysis those in the sample will have thought about their position more carefully than the rest of the population. However desirable it is that public policy should be based upon well informed and well thought opinion, I suspect that politicians would rather base it upon opinion as it is.

I.J. GOOD (*Virginia Polytechnic and State University*):

Dr. DeGroot referred to the situation where there are two or more statisticians who have specified predictive distributions for X. I think that theory is directly related to the problem of how a single statistician can improve his judgement, namely by comparing a number of procedures for specifying priors as if they were provided by several statisticians. In other words he can, so to speak, split his personality. One of the ways of seeing which statistician is better at predicting is by means of the logarithmic payoff function which I advocated in 1951 (Good, 1952). If the probability or probability density of the observed value of X (say x) is $p(x)$, the logarithmic payoff function is of the form $a+b$ log $p(x)$. This is one of the payoff functions that encourages the statistician to be honest, and when comparing two statisticians the gain of the first over the second is proportional to $\log[p_1(x)/p_2(x)]$ (in a self-explanatory notation). This has a further justification; we can imagine that there is a demiurge with perfect judgement whose probability (density) is $p_0(x)$. When comparing a statistician with the demiurge we could imagine that we were trying to find out which of the two was the demiurge. Then log $[p(x)/(p_0(x)]$ would be the weight of evidence in favour of the statistician's *being* the demiurge. We could imagine that we score *each* statistician in this way against the demiurge. Then the gain of statistician 1 over statistician 2 would be

$$\log\,[p_1(x)/p_0(x)] - \log\,[p_2(x)/p_0(x)] = \log\,[p_1(x)/p_2(x)]$$

so we don't need to know $p_0(x)$ for trying to decide which of the two is better. If there is a true probability density, then the expected advantage of 1 over 2 is

$$\int p_0(x)\log\,[p_1(x)/p_2(x)]dx$$

I have a comment concerning Dr. Box's comment. Dr. Box said that the observed ordinate $p(x)$ of the probability density should not be compared with the density at the mode, and so he asked about using the tail-area probability. If instead you compare $p(x)$ with the *average* value of the probability density then you would be using Warren Weaver's surprise index. I generalized Weaver's surprise index to a continuum of indexes in Good (1953, 1956) where I invented what has been called Rényi's generalized entropy. (Perhaps it should be attributed to Good). A special case of the generalization is $\int p(y)\log p(y)dy - \log p(x)$.

D.V. LINDLEY (*University College London*):

An alternative way of handling this problem is to suppose that the statistician is observed by a totally coherent person who takes the statistician's views and updates them in the light of experience with similar outcomes. This has been explored by Lindley, Tversky and Brown, (1979) . Equivalently, the statistician can think of his incoherent, natural self being monitored by a coherent person inside him. It is not obvious to me which approach is preferable but ours does appear to avoid the need for assumptions like (4.1), This conference has been dominated by technical papers and it is a real pleasure to welcome this thoughtful paper which tackles an important problem.

A. ZELLNER (*University of Chicago*):

In connection with DeGroot's suggested adaptive learning approach, consider two hypotheses regarding a parameter θ, namely $H_1{:}\theta = \theta_0$, a given value and $H_2{:}\theta \neq \theta_0$. If we have posterior probabilities for these hypotheses, p_1 and 1-p_1, the optimal (relative to a symmetric loss function) estimate of θ is $\hat{\theta} = p_1\theta_0 + (1-p_1)\bar{\theta}$, where $\bar{\theta}$ is the posterior mean of θ under H_2. $\hat{\theta}$ can be equivalently expressed as $\hat{\theta} = \theta_0 + (1-p_1)(\bar{\theta} - \theta_0)$ and it is seen that 1-p_1 is an "adjustment coefficient" that is data dependent. Similarly, when we consider two alternative models with posterior probabilities, p_1 and 1-p_1, the optimal point prediction is $\hat{y} = p_1\hat{y}_1 + (1-p_1)\hat{y}_2 = \hat{y}_1 + (1-p_1)(\hat{y}_2 - \hat{y}_1)$ where $\hat{y}_1$ and $\hat{y}_2$ are means of the predictive distributions for the two models. Again 1-p_1 appears as a data dependent adjustment coefficient. These traditional Bayesian procedures incorporate adaptive learning and thus there may be no need for an alternative learning model such as proposed by DeGroot.

REPLY TO THE DISCUSSION

M.H. DEGROOT (*Carnegie-Mellon University*):

I am grateful to all the discussants for their comments and their appreciation of the general problem that I am trying to attack in this paper. Both Dr. Dunsmore and Prof. Geisser comment on possible shortcomings and difficulties with the models that I have presented. As Dr. Dunsmore suggests, I should extend my models to cover shape misspecification and to include skewed distributions.

Prof. Geisser says that the learning process in my model doesn't provide for inspiration and is "better suited for plodding, dull, and unimaginative automatons" At first I thought that he was criticizing my model, but then I realized that he was actually pointing out that my model appropriately describes the learning process of most statisticians. More seriously, learning proceeds in my models neither too slowly nor too quickly, but at just the right rate, i.e., Bayesianly. If one wishes to allow for the "inspiration" of changing models based on the data, then these possible changes must be, and can be, incorporated into a supermodel.

I agree with these discussants—we do need better models. But I believe that the development of such models should go hand-in-hand with the necessary psychological modeling.

In answer to a question raised by Professor Geisser, precision misspecification is relevant, even if the statistician is only going to use the mean of his predictive distribution as his predicted value. Although the statistician's predicted values may be unbiased, he will find that they tend to be much closer to, or much farther from, the correct values than he anticipated. Incidentally, it would be nice if one fringe benefit of this work was to introduce colorful terms like "bias" and "unbiasedness" into Bayesian statistics and reclaim them from sampling theory statistics where they have been wasted on useless concepts.

As Professor Lindley suggests, I am sure that there are times when it can be helpful to suppose that an incoherent statistician has a shadowy coherent alter ego looking over his shoulder or a tiny coherent elf somewhere inside him struggling to emerge. But two aspects of my work should be emphasized: First, the statistician may be biased, but he is coherent. Second, an important purpose of the models is to reduce and ultimately eliminate the need for the statistician to carry on any dialogue with himself.

Professor Good also suggests that the statistician can split his personality and see which personality makes the best predictions. He should then, I suppose, adopt that personality (at least whenever he must make a prediction). Professor Good suggests the use of scoring rules to see which personality is doing best. One difficulty with the use of any particular scoring rule is that it must be assumed that the statistician's expected utility function is simply his expected total score over a sequence of predictions. But if the different personalities have different subjective probabilities, wouldn't they also have different utility functions? Again, I emphasize that one purpose of my models is to eliminate split personality, which is a step toward better mental health as well as better statistics.

Professor Zellner is correct in suggesting that the standard Bayesian methodology for choosing among different models may be adequate in describing the learning process in many situations. The essence of my models, however, is to carry over into future problems what we have learned in earlier problems about *how* to specify prior distributions. That idea seems to me to be new.

Professor Dawid makes several interesting and valid points. I do believe, however, that when the world tells the forecaster that his beliefs are out of touch with reality, the forecaster can recognize this message and make adjustments wholly within the Bayesian framework. He does not need to use the logic or methodology of significance testing, although a forecaster whose faith is weak would be tempted to do so. It is true that in order to make these adjustments, the forecaster must go to a hierarchical model with perhaps a large number, possibly even an infinite number, of levels. But if, as Professor Good states in his paper at this conference, the hyperparameters at the higher levels matter less and less, then he will have improved his forecasts.

Professor DuMouchel proposes a clever new model, and avoids the methodology of significance testing by carrying out a Bayesian test of his hypotheses at each stage. The model promises to be fairly comprehensive and clearly warrants further study, development, and application.

S.J. PRESS (*University of California, Riverside*):

The qualitative controlled feedback (QCF) data collection protocol is a procedure for collecting information of various kinds from a group; the information can be used and analyzed in a variety of ways. This broad base of applicability is one of the greatest assets of the approach. The procedure can be used for example, merely to collect arguments and justifications in favor of one policy or another that has been advocated. Group members can bring to bear arguments based upon information each of them has separately, and information they have generated together as a group, and they can also argue various positions on the basis of information they might not have originally, but later are exposed to, and they can evaluate it in a meaningful way. Group members may differ in the amounts of information they have available, the type of information they have available, and in their ability to verbalize arguments using this information. They will differ in their experience level, intellect, intuitive ability, and expertise. They will share however, a large base of intellect, rationality, and information. The variation in opinion, after several rounds of QCF, is in itself a measure of the uncertainty or lack of knowledge surrounding the situation. The results are therefore very meaningful even when consensus is not achieved. Many applications will involve no more than just a collection of arguments arrived at after several iterations of the QCF process. Such arguments may be useful for assessing risks and for evaluating a complicated situation. In other applications it may be useful to develop quantitative information about some important questions using opinion and arguments generated by the group. In these cases, the absolute answers may be of fundamental importance, or what may really be of interest is the change, over time, in the group's perception of the basic answers to the fundamental questions. In these kinds of applications it is useful to use the QCF procedure with a quantitative base. Finally, in still other applications, it may be useful to use a model, such as the one developed in the paper, for predicting the next round's quantitative outcome based upon earlier developed information. With these prefatory remarks I now turn to the thoughtful questions raised by participants at the Bayesian Conference.

Professor Bernardo raised the question of how one actually uses an inverted Wishart distribution in practice. An then, how does one use the more complicated $_pF_q$ generalized distribution discussed in the paper? This question is an important one from the point of view of practical applications of Bayesian methods in general, because the inverted Wishart distribution family is the one most often proposed as the family of natural conjugate priors that should be used for scale parameters. The inverted Wishart distribution of course has some problems associated with it, as I discussed in the paper, and these problems relate to there being some inherent constraints imposed on the parameters within the distribution, which the analyst may find undesirable. This problem was first pointed out by Rothenberg, (1963). The argument is also summarized in Press, (1972, page 233). Nevertheless, the parameters of the inverted Wishart distribution may be assessed by assessing quantiles of the marginal distributions, which of course are inverted gamma distributions. The quantiles are related to variances, medians, etc. Methods for assesssing quantiles of univariate distributions are by now well known; see for example, Schlaifer, (1961); Stael von Holstein, (1970); Winkler, (1967a and 1967b); Lindley, Tversky and Brown, (1979). Methods for assessing the correlation or covariance for higher dimensional distributions are currently being developed; see for example, Gokale and Press, (1979); Dawid, Dickey and Kadane, (1979); Kadane, Dickey, Winkler, Smith and Peters, (1978). There are also several computer routines that have been developed to assist the analyst in assessing the hyperparameters of prior distribution families such as the inverted Wishart (see Press, 1980 for a summary). Methods for assessing the parameters of generalized distributions involving generalized hypergeometric functions have not yet been developed. Such methods will depend upon development of the theory that relates to these distributions in terms of marginal and conditional distributions. Once these procedures are known, methods that have already been developed can be readily applied.

Professor Dunsmore was surprised that the Bayesian development of QCF appeared much later than the earlier development. The explanation is of course, that the earlier development emphasized the use of qualitative controlled feedback as a data gathering tool, while the Bayesian development imposed some distributional structure above and beyond that which was assumed earlier, and this structure permitted us to make posterior inferences about results that might be obtained on a later round of QCF that we are not able to carry out. Such an analysis, while interesting and useful in some applications, is not as generally applicable as is the basic data collection process inself. In terms of practical relevance of the procedure it should be understood that the QCF approach can be easily implimented in a real world context for one, or even several, questions of importance without any application of the modeling itself. The practicality of the modeling stems from the fact that in our limited experience involving an empirical application of the methodology (see Press, Ali, and Yang, 1979) we found that after three stages, the process had pretty much stablized. We anticipate that only two or three stages will be necessary for stablization of the process in more general situations as well. Thus, if there were two stages, and we wanted to predict a third, and we used precisely the same numbers that Professor Dunsmore suggests in his comment, the dimension of the **a** vector would be 45*. It is of course always possible, and often reasonable, to keep the dimension of the **a** vector small by the device of using only those reasons for the prediction of the next stage's response, which were given by large numbers of respondents, and deleting the remainder. In that case, the dimension of the coefficient vector would always remain quite manageable. I will not comment further on the coefficients varying with the number of stages, beyond my saying that the model assumes that they do not so vary, in order to maintain a parsimonious approach to the number of parameters in the problem.

Professor Dunsmore talked about a respondent who might reverse his position from that on an earlier round, at some point in the process. This should occur only when some new information has been introduced into the composite of explanations for respondent's answers. If such a turn-about were not based upon new information, other group members would be totally confused and disappointed by the apparent lack of rationality of the turnabout group member.

In ignoring the variable with the "largest" variance matrix we are merely ignoring one observation out of many, and the ignored observation is one which is known with decreasing precision as the sample size gets large. Such an approximation can clearly have little effect on the result. I share the implied concern in Professor Dunsmore's final plea, which is to "see this elegant statistical theory transformed into useful statistical practice". It is my fond hope that practitioners of statistical methodology in various areas will apply qualitative controlled feedback to practical problems.

Professor French has made some excellent points. His first is that the methodology presented in this paper is applicable to a situation in which there is a single decision maker who plans to use the opinions of the group to help him make his decision. Thus, the decision maker is in fact exterior to the group. This point will be made later in my comments to Professor Mouchart.

Next I must talk about "old Charlie" who has a gut feeling for winning projects. I was not pursuaded by this argument because I don't agree with Professor French that, "some of the best opinions are held without the holder being able to express why he holds them" This is the same kind of argument used by anti-Bayesians to show why the entire Bayesian approach is not useful. They claim that while Bayesians must use prior distributions to develop their analyses, most people cannot really quantify their judgments, and for this reason, it is usually impossible to assess a prior distribution.

* Because the process stabilized, we could take $R_2(1) = R_2(2) = 0$. So if $R_1(1) = 4$, $R_1(2) = 5$, the dimension of the **a** vector would be 45. Moreover, it may often be assumed that $\mathbf{a}_\alpha(j)$ does not vary with α and j, in which case the number of distinct elements in **a** that must be estimated is $\pi(n\text{-}1)$; so if $\pi = 5$, $n = 3$, we must estimate a 10-dimensional vector.

The limitations of these types of arguments have been elucidated on many occasions, so I will not repeat them here. In the application described in this paper it is of course necessary for people to introspect about their opinions, just as they would regarding a prior distribution. In this case, they must introspect to derive arguments for why they believe what they believe.

Professor French's final point deals with the question of how representative are the results developed in a qualitative controlled feedback data collection process. The answer is that the results obtained after several stages of QCF are representative of what would be obtained if a census were taken of the entire population, and QCF procedures where applied. Thus, public policy or any other kind of policy, can be formulated for a large population based upon careful reasoning of a "representative" subgroup.

Professor Geisser raised the very interesting question, of whether or not the QCF procedure could be misused by an individual who was trying to control the outcome of the procedure? The answer is of course, that the procedure could in fact be misused by a devious intermediary. He could manipulate the outcome by misrepresenting the composite that was fed back to the panal on each stage. This is mentioned briefly in section one of the final form of the paper. It is not anticipated, however, that in most applications the context would be one in which manipulation is likely. Of course, the effect of manipulation can always be minimized by using a small group of intermediaries, rather than a single individual, to accomplish the task of forming the composite of reasons.

With respect to the issue of "What's it all about Alfie", there are simple and straightforward answers. Suppose, as Professor Geisser suggests, the "mayor of Valencia wanted to resolve a pressing public policy issue say the building of a Bayesian Conference Center". First, I would commend the mayor on his good taste, assuming he was the one who exercised the foresighted leadership to suggest such a center. Next, I would propose that he use qualitative controlled feedback on questions posed before a panel of people appropriate to the political context of Valencia (a city council, a random sample of concerned citizens, etc.) In an "open discussion", the Mayor (with the help of other high ranking, very local, and strongly influential people) might very well bully Valencia into a decision that is really inappropriate for this city. Using QCF the decision would have to be made through careful reasoning and rational dialogue. A popular referendum shares some of the features of careful reasoning with QCF, but because certain very vocal and affluent groups advertise heavily to persuade people to their position, regardless of the common good or the rationality of the argument, such decisions are often inappropriate. The social psychological literature abounds with examples of how special interest groups tend to dominate such "open discussions" (for a summary, see, e.g. Press, 1978).

Professor Mouchart asked about the properties of the opinion pooling process proposed in this paper? The answer to this question derives from the context in which this procedure should be evaluated. The context was carefully detailed and discussed in an earlier paper; see Press, (1978). There, it was pointed out that our context is one in which we always assume there is a single decision maker who wants to take every group member's opinions into account, but he will make the final decision. This is the same context assumed by Kirkwood, (1972), and it avoids the conflicts and difficulties addressed by the Arrow "impossibility theorem". As a result of using this context, conventional decision theory applies to any decision made by a decision maker on the basis of QCF.

I would like to close by thanking the individuals who where kind enough to comment on the paper in an effort to clarify the nature of the process being discussed. I am also particularly grateful to Professor Dunsmore for his thoughtful suggestions for improving the format of the paper.

REFERENCES IN THE DISCUSSION

ARROW, K.J. (1963). *Social Choice and Individual Values*, New York: Wiley.

DAWID, A.P., DICKEY, J.M. KADANE, J.B.. (1979). Distribution theory and assessment methods for matrix *t* and multivariate *t* models. *Tech. Rep.* University College of Wales, Aberystwyth.

DEGROOT, M.H. (1979). Comments on Lindley *et. al. J. Roy. Statist. Soc. A.* **142**, 172-173.

FRENCH, S. (1978). *Consensus of opinion. Euro. J. Opl. Res.* (in press).

GOKHALE, D.V. and PRESS, S.J. (1979). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *Tech. Rep.* **58**, University of California, Riverside.

GOOD, I.J. (1952). Rational decisions. *J. Roy. Statist. Soc. B,* **14**, 107-114.

— (1953). The appropriate mathematical tools for describing and measuring uncertainty. Chapter 3 of *Uncertainty and Business Decisions*, 20-36. Liverpool: University Press.

— (1956). The surprise index for the multivariate normal distribution *Ann. Math. Statist.*, 1130-1135. Corrections, **1**. c. 28 (1957), 1055.

HARMAN, A.J. and PRESS, S.J. (1978). Assessing technological advancement using groups of experts. In *Formal Methods in Policy Formulation* (Bunn, D.W. and Thomas, H., eds.). 123-147. Basel: Birkhauser Verlag.

KADANE, J.B. and *et al* (1978). Interactive elicitation of opinion for a normal linear model. *Tech. Rep.* **150**, Carnegie-Mellon University.

KIRKWOOD, C.W. (1972). *Decision Analysis Incorporating Preferences of Groups.* Ph. D. Dissertation. Massachusets Institute University.

LINDLEY, D.V., TVERSKY, A. and BROWN, R.V. (1979). On the reconciliation of probability assessments (with discussion) *J. Roy. Statist. Soc. A* **142**, 146-180.

LUCE, R.D. and RAIFFA, H. (1958). *Games and Decisions.* New York: Wiley.

MURPHY, A.H. and WINKLER, R.L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *App. Statist.* **26**, 41-47.

PATTERNAIK, P.K. (1978). *Strategy and Group Choice.* Amsterdam: North-Holland.

PRESS, S.J. (1972). *Applied Multivariate Analysis.* New York: Holt, Rinehart and Winston, Inc.

— (1979). Qualitative controlled feedback for forming group judgments and making decisions. *J. Amer. Statist. Assoc.* **73**, 526-535.

— (1980). Bayesian computer programs. In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys.* (Zellner, A., ed.) 429-442. Amsterdam: North Holland.

PRESS, S.J., ALI, M.W. and YANG, E. (1979). An empirical study of a new method for forming group judgments: Qualitative controlled feedback. *Technological Forecasting and Social Change*, **15**, 171-189.

ROTHENBERG, R.J. (1963). A Bayesian analysis of simultaneous equation systems. *Tech. Rep.* **6315** Rotterdam: Econometric Institute.

SCHLAIFER, R. (1961). *Introduction to Statistics for Business Decisions.* New York: McGraw-Hill.

WINKLER, R.L. (1967a). The assessment of prior distributions in Bayesian analysis. *J. Amer. Statist. Assoc*, **62**, 776-800.

— (1967b). The quantification of judgment: Some methodological suggestions. *J. Amer. Statist. Assoc.* **62**, 1105-1120.

Trab. Estadist. (1980)
Vol. **31**, pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 431-468 Valencia, Spain: University Press.

10. Predictive sample reuse

INVITED PAPER

GEISSER, S. (*University of Minnesota*)

Predictive sample reuse techniques for censored data

DISCUSSANTS

GUTTMAN, I. (*University of Toronto*)
PRESS, S.J. (*University of California, Riverside*)

REPLY TO THE DISCUSSION

Predictive sample reuse techniques for censored data

S. GEISSER

University of Minnesota

SUMMARY

Predictive sample reuse methods usually applied in low structure aparametric paradigms are shown to be useful in certain high structure situations when conjoined with a Bayesian approach. Particular attention is focused on the incomplete data situation for which two alternative sample reuse approaches are devised. The first involves differential weighting and the second a recursive sample reuse algorithm. There are applied to censored exponential survival data. The algorithmic approach appears to be preferable from both a computational and modelling viewpoint.

Keywords: BAYES, CONDITIONAL PREDICTIVE FUNCTION, CENSORED DATA, DISCREPANCY MEASURE, FUNGIBLE, EMPIRICAL BAYES, MAXIMUM LIKELIHOOD, METHOD OF MOMENTS, PREDICTIVE DISTRIBUTION, PREDICTIVE SAMPLE REUSE, PSEUDO-OBSERVATIONS, SAMPLE REUSE ALGORITHM.

1. PREDICTIVE SAMPLE REUSE

The predictive sample reuse (PSR) method was presented in a variety of detailed forms, Geisser (1974, 1975a), Stone (1974). Here we shall delineate it in a very simple manner appropriate to the particular applications that flow from it under discussion in later sections.

Suppose we have a set of observations $\mathbf{x}^{(N)} = (x_1,\dots,x_N)$ and we are interested in predicting a future observation from the process generating observations of this kind. We further assume a predictive function used to forecast a potentially observable value,

$$x_{N+1} = f(\mathbf{x}^{(N)}, \alpha) \tag{1.1}$$

where α is defined as some unknown constant or set of such unknowns whose domain is Ω. Next we define a discrepancy function

$$D(\alpha) = D(d_1,\ldots,d_N,\Sigma) \tag{1.2}$$

where $d_j = d(x_j,f_j)$ represents a discrepancy between the observed value x_j and $f_j = f_j(\mathbf{x}_j^{(N-1)}, \alpha)$ which is defined as in (1.1) except that x_j has been deleted from f and $\Sigma = \Sigma(\alpha, \mathbf{x}^{(N)})$ represents some scheme of weighting the various d_j singly or jointly. For example

$$D(\alpha) = \Sigma_{j=1}^{N} a_j(\alpha) d(x_j, f_j) \tag{1.3}$$

where $a_j(\alpha)$ is the weight assigned to the j^{th} discrepancy or

$$D(\alpha) = \mathbf{d}'\Sigma\mathbf{d} \tag{1.4}$$

for $\mathbf{d}' = (d_1,\ldots,d_N)$ would be two such schemes. In most cases fungible[1] data would lead to $a_j(\alpha) = N^{-1}$ or $\Sigma = N^{-1}\mathrm{I}$. Then $D(\alpha)$ is minimized for values of α restricted to Ω which we assume yields a unique value $\hat{\alpha}$. This leads to the predictor

$$\hat{\mathbf{x}}_{N+1} = f(x^{(N)}, \hat{\alpha}) = \hat{f}. \tag{1.5}$$

For a more detailed exposition of the method involving multiple observational omissions and various schemata of omission, as well as applications, see Geisser (1974, 1975a, 1976).

In applying this method to survival or realiability data, it is quickly apparent that an inherent deficiency exists. The method as stated depends on the full knowledge of the sample values. But for this type of problem quite often our knowledge for a portion of the sample is restricted by the fact that the observations were censored at particular values. In order to remedy this lack of knowledge of fully observed values we introduce pseudo-observations. They depend on α and are determined from defined conditional predictive functions. Two procedures utilizing a pseudo-observation approach are presented. The first proposal substitutes the pseudo-observations into the discrepancy measure prior to minimization. This leads rather naturally to considering schemes whereby the censored observations are weighted differently than uncensored ones as opposed to previous applications where $a_j(\alpha) = 1$ on the basis that the data were inherently fungible. Of course, there could arise situations where a sample of uncensored observations may require different weights because of a decision as to their treatment or a model for their generation. Here, even though we start with a scheme that treats the observations fungibly, the approach of fitting the censored observations into the predictive sample reuse framework naturally induces consideration of differential weighting schemes.

A second proposal involves the substitution of the pseudo-observations into the solutions as if all the values were fully observed and solving the requisite algorithm. Let $\mathbf{x}^{(d)} = (x_1,\ldots,x_d)$ and $\mathbf{x}^* = (x_{d+1},\ldots,x_N)$ represent respectively the completely and partially observed data sets with the understanding that the observable x_j for $j > d$ represents incomplete information of some kind on an observable entity, or when appropriate, a realization of the random variable X_j. Let $\mathbf{y} = (y_{d+1},\ldots,y_N)$ represent the set of values which would have been observed but were partially observed as $\mathbf{x}^*$; i.e. the fully realized value of X_j would have been y_j,but we were only able to record the partially observed value x_j, $j > d$. We then compute a complete solution for α, say

$$\tilde{\alpha} = \tilde{\alpha}(\mathbf{x}^{(d)},\mathbf{y}) \tag{1.6}$$

in the usual fashion, as in the fully observed case, but as a function of $\mathbf{y}$. But we need values for y_j the components of $\mathbf{y}$. We now assume a conditional predictive function for the components of $\mathbf{y}$,

$$y_j = \hat{x}_j'(\mathbf{x}^{(d)},\mathbf{x}^*, \alpha) = x_j'(\alpha); \quad j > d. \tag{1.7}$$

Now let $\mathbf{x}^*(\alpha)$ represent the set of values inserted for $\mathbf{y}$; i.e. for each component y_j we insert $x_j'(\alpha)$ in (1.6). Lastly we then have the algorithm

$$\alpha = \tilde{\alpha}(\mathbf{x}^{(d)}, \mathbf{x}^*(\alpha)) \tag{1.8}$$

which needs to be solved for α. Call the solution $\hat{\alpha}$ and one then uses this either to predict a future observation conditionally or unconditionally.

2. AN APPLICATION--UNCENSORED CASE

The application of these ideas for forecasting in a particular survival or reliability data situation will be presented where the predictive sample reuse technique is used in partial conjunction with a Bayesian approach. Initially we shall assume the entire fine structure of an exponential survival distribution *cum* gamma prior distribution on the exponential parameter. Subsequently

[1] We use the term fungible to extend the notion of exchangeable to data that are not necessarily a realization of a random set of variables. For random variables the terms are equivalent. The extension, though ill defined, conveys an attitude that one could take towards observable data for which it is inappropriate to assume that they were necessarily generated by a random process.

the predictive distribution of a future observation from the process is obtained. In the gamma prior we essentially assume one of the hyperparameters known (or guessed) and the other unknown. An estimate for the latter is produced by the predictive sample reuse method essentially as a by-product of deriving a point predictor. The question of censored data, where ambiguity exists in the execution of the predictive sample reuse method is treated in the next section and tentatively resolved by the ploy of pseudo-observations that are supplied from a partial Bayesian or other structure.

The utilization of the approximate predictive distribution;i.e. with one hyperparameter estimated, as a forecasting tool is valid to the extent of the appropriateness of the fine structure assumptions with uncertainty commensurate with the roughness of the approximation. On the other hand the predictor itself may be useful considerably beyond the bounds of the initial structure assumed in that it may be robust as a point predictor for a variety of possible structures. Further it may be most useful in a low structure situation, where any specific distributional assumptions are fraught with peril.

Suppose we have a random sample $X_1,\ldots,X_N$ on an exponential random variable X whose density is

$$f\,(x\,|\,\mu) = \mu e^{-\mu x}; \quad \mu > 0, \quad x > 0. \tag{2.1}$$

If our prior objective or subjective information is subsumed in a prior density for μ,

$$p\,(\mu) \propto \mu^{\delta-1}e^{-\gamma\mu}; \quad \gamma > 0, \delta > 0 \tag{2.2}$$

and we are interested in predicting a value x_{N+1} for the random future observation X_{N+1} given the previous N observations $\mathbf{x}^{(N)}$; say, then the predictive density for X_{N+1} is easily calculated to be, for $x_{N+1} > 0$,

$$\begin{aligned} f\,(x_{N+1}|\mathbf{x}^{(N)}) &= \int p\,(\mu|\mathbf{x}^{(N)})f\,(x_{n+1}|\mu)\,d\mu \\ &= (N+\delta)\,(N\bar{x}+\gamma)^{N+\delta}/(N\bar{x}+\gamma+x_{N+1})^{N+\delta+1} \end{aligned} \tag{2.3}$$

where $\bar{x}$ is the sample mean and $p(\mu\,|\,x^{(N)})$ is the posterior density of μ given the previous N observations $x^{(N)}$. Hence our forecast about X_{N+1} involves the hyperparameters γ and δ which enter the problem via the distribution of the parameter μ. Before any observations are taken one can also find the predictive (marginal) density of the generic variable X, namely

$$f(x) = \int f(x\,|\,\mu)\,p(\mu)\,d\mu = \delta\gamma^{\delta}/(\gamma+x)^{\delta+1}; \quad x > 0. \tag{2.4}$$

Hence it is convenient and perhaps more appropriate to think about these hyperparameters in terms of predicting X before any observations are taken rather than in how they modulate the assumed prior distribution of μ. Therefore, prior to the sample, we have

$$\begin{aligned} \mathrm{E}(X) &= \gamma/(\delta-1) = g \\ \mathrm{Var}\,(X) &= \delta\gamma^2/(\delta\text{-}2)(\delta\text{-}1)^2 = g^2(1+\alpha)/(1\text{-}\alpha) \end{aligned} \tag{2.5}$$

where $\alpha = (\delta\text{-}1)^{-1}$.

Clearly Var(X) exists for $0<\alpha<1$, and E(X) exists for $\alpha>0$ while the distribution exists for all $\alpha \notin [-1,0]$. Hence if one could frame his prior opinions about the potentially observable values of X in terms of its expectation and variance then one can easily execute the whole predictive process by solving for the appropriate values δ and γ from (2.5) and substituting them in (2.3).

It is to be noted that (2.3) and (2.4) were obtained from (2.1) and (2.2). However, for the predictivist who would prefer to start from (2.1) and (2.4) in terms of convenience of framing his predictions this is somewhat awkward. Interestingly enough in this case starting with $f(x\,|\,\mu)$ and $f(x)$ is sufficient to obtain $p(\mu)$ and $f(x_{N+1}|\bar{x})$ which is a more logical and appealing approach for the predictivist. This is true here because $f(x)$ is the unique Laplace transform of $\mu^{-1}p(\mu)$.

Now as we mentioned previously making all of these assumptions yields the requisite information for making probability statements about a future value provided that one has specified values for g and α. However while one may often be willing to hazard a guess at g, one may be far less willing to specify a value for α.

We now shall apply the predictive sample reuse method in order that the data itself should yield a value for α once g has been assumed.

If we had already observed $\mathbf{X}^{(N)} = \mathbf{x}^{(N)}$ and wished to predict a future value for X_{N+1}, we could use the posterior expectation of X_{N+1} obtained from the predictive density given by (2.3). This is easily calculated to be

$$E(X_{N+1}) = (N\bar{x}+\gamma)/(N+\delta-1) = (\alpha N\bar{x}+g)/(\alpha N+1) = f. \tag{2.6}$$

Note that when $\delta \rightarrow 1$ and $\gamma \rightarrow 0$, we obtain the usual predictor $\bar{x}$.

In terms of the predictive sample reuse method, Geisser (1975), equation (2.6) may be utilized as a predictive function. In order to supply a value for α we apply the method using one-at-a-time omissions and a squared discrepancy as follows: The average squared discrepancy is

$$D(\alpha) = N^{-1}\Sigma(f_i - x_i)^2 = N^{-1}\Sigma\left(\frac{\alpha(N\text{-}1)\bar{x}_i + g}{\alpha(N\text{-}1) + 1} - x_i\right)^2 \qquad (2.7)$$

where f_i and $\bar{x}_i$ are defined respectively as the predictive function and the sample average with x_i omitted. In order to find a suitable α, we minimize $D(\alpha)$ with respect to α for $\alpha \geq 0$. (Note again that for the density given by (2.4), Var (X) exists only for $0 < \alpha < 1$, although the distribution for X exists for $\delta > 0$ and hence for all $\alpha \notin [-1,0]$. Nevertheless we shall not restrict ourselves to $\alpha > 0$ although this is essentially the range on α for which the prior mean exists), but also include $\alpha = 0$, a value, which is possible when γ is a function of α and $\alpha \rightarrow g$ as $\alpha \rightarrow 0$.)

We can easily evaluate

$$D(\alpha) = [(N\text{-}1)s^2(\alpha N+1)^2 + N(g\text{-}\bar{x})^2] \,/\, N[\alpha(N\text{-}1) + 1]^2, \qquad (2.8)$$

where $s^2 = (N\text{-}1)^{-1}\,\Sigma_{i=1}^{N}\,(x_i\text{-}\bar{x})^2$. Taking the derivative with respect to α and setting this equal to zero yields the solution

$$\begin{aligned} \hat{\alpha} &= (t^2\text{-}1)/N && \text{for } t^2 > 1 \\ \hat{\alpha} &= 0 && \text{if } t^2 \leq 1 \end{aligned} \qquad (2.9)$$

where $t^2 = N(g\text{-}\bar{x})^2/s^2$. Hence this yields the predictor

$$\begin{aligned} f(\hat{\alpha}) &= \hat{f} = [(t^2\text{-}1)\bar{x}+g]/t^2 && \text{if } t^2 > 1 \\ f(\hat{\alpha}) &= g && \text{if } t^2 \leq 1 \end{aligned} \qquad (2.10)$$

Of course for the strict Bayesian the use of $\hat{\alpha}$ and its derived value $\hat{\delta}$ contradicts the fundamental canon of Bayesianism that the prior hyperparameters should not depend on the data. However it should serve as an approximate solution to the problem in the sense that the unknown hyperparameter δ is replaced by $\hat{\delta}$ if $\hat{\alpha} > 0$ in (2.3), given the high structure assumptions. This problem and method for solution was first proposed by Geisser (1975b) with further commentary, Geisser (1976, 1980).

It may also be mentioned that the predictor $\hat{f}$ can also be conceived as totally independent of the Bayesian process and the likelihood when obtained from this approach in the sense that we have merely chosen f as a point predictor for X_{N+1} and have ascertained $\hat{f}$ by a squared discrepancy measure. We also note that the predictive function f is basically a linear combination of the mean $\bar{x}$ and the prior guess g with weights αN and 1. There are undoubtedly other models that can lead to forecasting the next observation as linear combinations of a prior mean and the sample mean when the predictive expectation of a future observation is utilized. In this regard then one could define a predictive function that is a linear combination of the mean and a guessed value g

$$f^* = \alpha^*\bar{x} + (1\text{-}\alpha^*)g, \quad 0 \leq \alpha^* \leq 1 \qquad (2.11)$$

This yields, for squared discrepancy and one-at-a-time omissions, Geisser (1975a),

$$\begin{aligned} \alpha^* &= (t^2\text{-}1)/[t^2+(N\text{-}1)^{-1}] && \text{for } t^2 > 1, \\ &= 0, && \text{for } t^2 \leq 1 \end{aligned} \qquad (2.12)$$

Hence

$$\begin{aligned} \hat{f}^* &= [(t^2\text{-}1)\bar{x} + N(N\text{-}1)^{-1}g]/[t^2+(N\text{-}1)^{-1}], && \text{for } t^2 \geq 1 \\ &= g && \text{if } t^2 < 1 \end{aligned} \qquad (2.13)$$

Clearly $\alpha^* = \alpha N/(\alpha N+1)$ for $\alpha \geq 0$ in terms of the transformed predictive function. On the other hand $\hat{\alpha}^* < \hat{\alpha}N/(\hat{\alpha}N+1)$, for $t^2 > 1$, the estimation procedure not being invariant under such a transformation. However they will be quite close as they are asymptotically equivalent for large N. Comparison of $\hat{f}$ with $\hat{f}^*$ reveals they also converge for large N, but slightly more weight is attached to $\bar{x}$ in $\hat{f}$ than in $\hat{f}^*$.

In summary then, in the assumed presence of the high initial structure f should be preferable, but for robustness to other structures leading approximately to the aforementioned linear combination, f^* may be preferable. In any event the difference is negligible for large N. In the absence of any distributional assumptions both predictors are viable methods for having something to say about the prediction of future observations.

3. CENSORED DATA

In many cases especially in survival or reliabilitity studies the experiment is usually terminated before all of the subjects or units have expired or failed. Suppose the experiment is such that for d of the observations, failure times are recorded as $x_1, \ldots, x_d$, while the remaining N-d observations have survived but were censored at values $x_{d+1}, \ldots, x_N$. Hence

$$L(\mu) = \Pi_{i=1}^{d} f(x_i|\mu)\, \Pi_{i=d+1}^{N} [1-F(x_i|\mu)]$$

where $F(x_i|\mu)$ is the distribution function of X_i. For the exponential case, clearly

$$L(\mu) \propto \mu^d e^{-\mu[dx_d + (N-d)x_{N-d}]} \tag{3.1}$$

where $\bar{x}_d = d^{-1} \Sigma_1^d x_i$ and $\bar{x}_{N-d} = (N-d)^{-1} \Sigma_{i=1}^{N-d} x_{d+i}$. From (3.1) and (2.2) we can obtain first the posterior density of μ and then, as previously, the predictive density for a future observation X_{N+1},

$$\begin{aligned} &f(x_{N+1}|\mathbf{x}^{(d)},\mathbf{x}^{(N-d)}) \\ &\quad = (d+\delta)(d\bar{x}_d + (N-d)\bar{x}_{N-d}+\gamma)^{d+\delta}/(d\bar{x}_d + (N-d)\bar{x}_{N-d}+\gamma+x_{N+1})^{d+\delta+1} \end{aligned} \tag{3.2}$$

where $\mathbf{x}^{(d)}$ represents the observations whose failure times are recorded and $\mathbf{x}^{(N-d)}$ the censored observations. Further the predictive expectation, to be used as the predictive function, is

$$\begin{aligned} E(X_{N+1}) &= [d\bar{x}_d + (N-d)\bar{x}_{N-d} + \gamma]/(d + \delta - 1) \\ &= [(d\bar{x}_d + (N-d)\bar{x}_{N-d})\alpha + g]/(\alpha d+1) = f. \end{aligned} \tag{3.3}$$

Note that for $\delta \to 1$ and $\gamma \to 0$ we obtain the usual predictor

$$\bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d}.$$

Due to censoring there is difficulty in appropriately executing the predictive sample reuse method. One tentative solution is to generate N-d pseudo-observations having values x'_{d+i}, $i=1,\dots,$ N-d, say. These are the presumed failure times for the censored observations $x_{d+1},\dots,x_N$. We shall take as the pseudo value x'_{d+i}, the expectation of the predictive distribution of X_{d+i} given $X_{d+i} > x_{d+i}$, the censored value. More precisely the likelihood in (3.1) is used but with x_{d+i} omitted; i.e., based on all the observations but x_{d+i}. This is then combined with the prior density of μ whence the posterior density of μ is obtained and subsequently the predictive density of X_{d+i} computed. From this we then compute the conditional density of X_{d+i} given $X_{d+i} > \mathbf{x}_{d+i}$

$$f(x|X_{d+i} > \mathbf{x}_{d+i}) = \frac{(d+\delta)(d\bar{x}_d + (N-d)\bar{x}_{N-d}+\gamma)^{d+\delta}}{(d\bar{x}_d + (N-d)\bar{x}_{N-d}+\gamma+x-x_{d+i})^{d+\delta+1}} \tag{3.4}$$

Further computation yields

$$\begin{aligned} E(X_{d+i}|X_{d+i} > x_{d+i}) &= [(d+\delta-1)x_{d+i} + d\bar{x}_d + (N-d)\bar{x}_{N-d}+\gamma]/(d+\delta-1) \\ &= x_{d+i} + \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d})\alpha + g}{\alpha d + 1} = x'_{d+i}, \end{aligned} \tag{3.5}$$

and

$$\mathrm{Var}(X_{d+i}|X_{d+i} > x_{d+i}) = \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d}+\gamma)^2(d+\delta)}{(d+\delta-1)^2(d+\delta-2)} = \frac{d+\delta}{d+\delta-2} f^2 \tag{3.6}$$

the latter being independent of i.

Now in executing the sample reuse method with predictive function given by (3.3) using the actual observations $x_1,\dots,x_d$ and the pseudo observations $x'_{d+1},\dots,x'_N$ given by (3.5) it seems sensible to give the pseudo-observations a weight that differs from that assigned to the uncensored observations in contradistinction to an unweighted and consequently inadequate solution, Geisser (1975b). We note that

$$\mathrm{Var}(X_i|\mu) = \mu^{-2} \qquad \text{for } i=1,\dots,d. \tag{3.7}$$

Since μ is unknown we shall compute

$$E_\mu[\mathrm{Var}(X_i|\mu)] = E_\mu[\mu^{-2}] \tag{3.8}$$

over the posterior distribution of μ. This results in

$$E_\mu(\mu^{-2}) = \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d}+\gamma)^2}{(d+\delta-1)\,(d+\delta-2)} = \frac{d+\delta-1}{d+\delta-2} f^2 \tag{3.9}$$

where f is as defined in (3.3).

We can define a weighted discrepancy for $d > 1$, N-$d > 1$ as follows:

$$\begin{aligned} D(\alpha) = {} & E_\mu^{-1}(\mu^{-2})\, \Sigma_{j=1}^{d}\left(\frac{[(d-1)\bar{x}_{d,j} + (N-d)\bar{x}_{N-d}]\alpha + g}{\alpha d+1} - x_j\right)^2 \\ & + [\mathrm{Var}(X|X > x_{d+i})]^{-1}\, \Sigma_{k=d+1}^{N}\left(\frac{[d\bar{x}_d + (N-1-d)\bar{x}_{N-d,k}]\alpha + g}{\alpha d+1} - x'_k\right)^2 \end{aligned} \tag{3.10}$$

where $x_{d,j}$ and $x_{N-d,k}$ are respectively the sample means of d-1 uncensored observations omitting x_j and the mean of N-1-d censored observations omitting x_k.

After some algebraic manipulation we obtain

$$\mathrm{D}(\alpha) = \frac{(d\text{-}1)s_d^2(\alpha d+1)^3 + d(g\text{-}\bar{x}_d+\alpha(\mathrm{N}\text{-}d)\bar{x}_{N-d})^2(\alpha d+1)}{|\alpha(d\text{-}1)+1][(d\bar{x}_d+(\mathrm{N}\text{-}d)\bar{x}_{N-d})\alpha+g]^2}$$
$$+ \frac{|\alpha(d+1)+1][\alpha(d\text{-}1)+1]}{[(d\bar{x}_d+(\mathrm{N}\text{-}d)\bar{x}_{-d})\alpha+g]^2}\,\Sigma_{j=d+1}^{N}x_j^2. \qquad (3.11)$$

The solution then for α is obtained by differentiating (3.11) with respect to α and setting it equal to zero. This will result in a polynomial in α, whose roots are stationary points. After discarding negative and complex roots, the positive roots α, say, need be compared with D(0) and D(∞) to ascertain the global minumum for $\alpha \geq 0$.

For $d=1$ and $\mathrm{N}>2$ only the second term in (3.11) obtains and formal minimization in this case yields $\alpha=\infty$, so that $f=\mathrm{N}\bar{x}$, the usual predictor in this case.

For $d>1$ and $\mathrm{N}=d+1$ only the first term in (3.11) obtains. Minimization then follows in the same manner as in the discussion for $d>1$ and $\mathrm{N}\text{-}d>1$.

It is to be noted that in the weighting we merely used terms that reflected variation. Perhaps a more appropriate weighting scheme would also include covariation among those values that are correlated. As a step in this direction we can take cognizance of the covariance among the pseudo-observations.

A simple calculation reveals that the joint predictive density of X_{d+i} and X_{d+j} $i \neq j = 1, \ldots, \mathrm{N}\text{-}d$ conditional on $\mathrm{X}_{d+i} > x_{d+i}$ and $\mathrm{X}_{d+j} > x_{d+i}$ is

$$f(z,w\,|\,\mathrm{X}_{d+i}>x_{d+i},\mathrm{X}_{d+j}>x_{d+j}) = \frac{(d+\delta)(d+\delta+1)(d\bar{x}_d+(\mathrm{N}\text{-}d)\bar{x}_{N-d}+\gamma)^{d+\delta}}{(d\bar{x}_d+(\mathrm{N}\text{-}d)\bar{x}_{N-d}+z\text{-}x_{d+i}+w\text{-}x_{d+j})^{d+\delta+2}} \qquad (3.12)$$

whence we calculate

$$\mathrm{Cov}(\mathrm{X}_{d+i}\mathrm{X}_{d+j}\,|\,\mathrm{X}_{d+i}>\mathrm{X}_{d+j}>x_{d+j}) = (d+\delta)^{-1}\,\mathrm{Var}(\mathrm{X}_{d+i}\,|\,\mathrm{X}_{d+i}>x_{d+i}), \qquad (3.13)$$
$$\text{for } i \neq j,\ i, j = 1, \ldots, \mathrm{N}\text{-}d$$

Use of this alters the second term in (3.11) to

$$\frac{[\alpha(d+1)+1][\alpha(d\text{-}1)+1]}{(\alpha\mathrm{N}+1)(\alpha d+1)[(d\bar{x}_d+(\mathrm{N}\text{-}d)\bar{x}_{N-d})\alpha+g]^2}$$
$$\times\ \ [\alpha(\mathrm{N}\text{-}1)+1)\,\Sigma_{j=1}^{N-D}x_{d+j}^2\text{-}2\alpha\Sigma_{i>j}^{N-D}x_ix_j] \qquad (3.14)$$

When, as is often the case, all of the N-d observations are censored at the same value, say x_o, then (3.14) simplifies to

$$\frac{[\alpha(d+1)+1]^2[\alpha(d\text{-}1)+1](\mathrm{N}\text{-}d)x_o^2}{(\alpha\mathrm{N}+1)\,[(dx_d+(\mathrm{N}\text{-}d)x_o)\alpha+g]^2} \qquad (3.15)$$

This term is then $[\alpha(d+1)+1]/[\alpha\mathrm{N}+1]$ times the second term in (3.11), indicating roughly the disminished effect of the contribution of the portion of D(α) involving the pseudo-observations by taking into account their covariance structure. Of course this further complicates arriving at a solution for α and it is not clear just how significant the resulting improvement would be.

The most complex weighting scheme would also attempt to take into account covariation between uncensored observations and pseudo-observations. Now for $i=1, \ldots, d$, $j=d+1,\ldots,\mathrm{N}$; $\mathrm{X}_j' = \mathrm{X}_j + (\mathrm{N}\alpha\bar{\mathrm{X}}+g)/(\alpha d+1)$

$$\mathrm{Cov}(\mathrm{X}_i,\mathrm{X}_j'|\mu) = \frac{\alpha}{\alpha d+1}\,\mathrm{V}(\mathrm{X}_i|\mu) = \frac{\alpha\mu^{-2}}{\alpha d+1}. \qquad (3.16)$$

Again using (3.9) we find that

$$\mathrm{E}_\mu[\mathrm{Cov}(\mathrm{X}_i,\mathrm{X}_j'|\mu] = f^2/(d+\delta\text{-}2). \qquad (3.17)$$

Hence we may use as a weighting matrix the inverse of the NxN partitioned matrix

$$V = f^2/(\mathrm{d}+\delta\text{-}2)\ \left(\begin{array}{cc} \mathrm{d} & \mathrm{N\text{-}d} \\ (\mathrm{d}+\delta\text{-}2)\mathrm{I} & \mathrm{J}_{12} \\ \mathrm{J}_{21} & (d+\delta\text{-}1)\mathrm{I}+\mathrm{J}_{22} \end{array}\right)\begin{array}{c} \\ d \\ \mathrm{N}\text{-}d \end{array} \qquad (3.18)$$

where J_{ij} is a matrix all of whose entries are unity. The inverse of V can readily be displayed by letting $\mathrm{U} = f^2(\alpha d+1)[\alpha(d\text{-}1)+1]^{-1}\mathrm{V}^{-1}$ with partitions similar to V so that

$$\mathrm{U}_{11} = \mathrm{I} + \frac{(\mathrm{N}\text{-}d)\mathrm{J}_{11}}{(d+\delta\text{-}1)(\mathrm{N}+\delta\text{-}1)\text{-}d(\mathrm{N}\text{-}d)}$$

$$\mathrm{U}_{ij} = \frac{\text{-}(d+\delta\text{-}1)\mathrm{J}_{ij}}{(d+\delta\text{-}1)(\mathrm{N}+\delta\text{-}1)\text{-}d(\mathrm{N}\text{-}d)},\ \text{for } i \neq j \qquad (3.19)$$

$$U_{22} = I \frac{(\delta-1)J_{22}}{(d+\delta-1)^2+(\delta-1)(N-d)}$$

Now for $d > 1$ and $N-d > 1$, let

$$\begin{aligned} \triangle_j &= f_j - x_j && \text{for } j = 1,\dots,d \\ &= f_j - x_j' && \text{for } j = d+1,\dots,N \end{aligned} \qquad (3.20)$$

where again f_j is the predictive expectation f omitting the j^{th} observation. Further, letting $\triangle' = (\triangle_1,\dots,\triangle_N)$ we can now define

$$D(\alpha) = \triangle' V^{-1} \triangle$$

and minimize it for $\alpha > 0$. Again evaluation of $D(\alpha)$ leads to rather complicated algebra which we shall omit.

Once a solution $\hat{\alpha}$ is rendered we can convert it to obtain the approximate predictive distribution of a future observation or just use $\hat{f}$ as a point predictor.

For the second kind of predictive function

$$f^* = \alpha^*(\bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d}) + (1-\alpha^*)\, g = \alpha^* h + (1-\alpha^*)\, g \qquad (3.21)$$

which does not lean as much on the previous high structure assumptions, we use as pseudo-observations

$$x'_{d+i} = x_{d+i} + \bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d} = x_{d+i} + h. \qquad (3.22)$$

This is akin to frequentist prediction since using x'_{d+i}, $i = 1,\dots,N-d$ as actual observations in conjunction with $x_1,\dots,x_d$ preserves the frequentist predictor, $\bar{x}_d + d^{-1}(N-d)\bar{x}_{N-d}$, as this is the average of both uncensored values and pseudo-observations. Now (3.22) can also be obtained by letting $\delta \to 1$ and $\gamma \to 0$ in (3.5).

Here the simplest weighted squared discrepancy measure neglecting covariation but not variances is

$$D(\alpha^*) \propto \Sigma_{i=1}^{d} (f_j^* - x_j)^2 + \frac{d}{d+1} \Sigma_{j=d+1}^{N} (f_j^* - x_j)^2 \qquad (3.23)$$

where f_j^* is f^* as in (3.21) but with x_j omitted. The weighting here is again closer to a frequentist approach although it also can be obtained from (3.6) and (3.9) by letting $\delta \to 1$. Let $f_j^* = \alpha^* h_j + (1-\alpha^*)g$ so that

$$\begin{aligned} h_j &= (d-1)^{-1}(d\bar{x}_d + (N-d)\bar{x}_{N-D} - x_j) && \text{for } j = 1,\dots,d \\ &= \bar{x}_d + d^{-1}[(N-d)\bar{x}_{N-d} - x_j] && \text{for } j = d+1,\dots,N \end{aligned} \qquad (3.24)$$

then the minimization of $D(\alpha^*)$ with respect to α^* yields

$$\begin{aligned} \hat{\alpha}^* &= \frac{\Sigma_{j=1}^{d} (h_j-g)(x_j-g) + d\,(d+1)^{-1} \Sigma_{j=d+1}^{N} (h_j-g)(x_j'-g)}{\Sigma_{j=1}^{d} (h_j-g)^2 + d(d+1)^{-1} \Sigma_{j=d+1}^{N} (h_j-g)^2} && \text{for } 0 \le \hat{\alpha}^* \le 1 \\ &= 1 && \text{for } \hat{\alpha}^* > 1 \\ &= 0 && \text{for } \hat{\alpha}^* < 0. \end{aligned} \qquad (3.25)$$

If one uses a scheme with no weighting at all then

$$\begin{aligned} \hat{\alpha}^* &= \frac{(d-1)N(h-g)^2 + (h-g)d^{-1}(N-d)\bar{x}_{N-d} - d^{-1}(N-d)^2\bar{x}_{N-d}^2 - (d-1)s_d^2 - (d-1)d^{-1}\Sigma_{d+1}^{N} x_j^2}{(d^2-1)(h-g)^2 + 2(h-g)(N-d)\bar{x}_{N-d} + (d-1)^{-1}d^{-1}\bar{x}_{N-d}^2 + s_d^2 + (d-1)d^{-2}\,\Sigma_{d+1}^{N} x_j^2} \\ &= 0 \qquad \text{if } \hat{\alpha}^* \le 0 \\ &= 1 \qquad \text{if } \hat{\alpha} \le 1. \end{aligned} \qquad (3.26)$$

A slightly different solution can be obtained by altering the function h. Previously h was defined as the sum of all the observations censored and uncensored, divided by the number of uncensored observations. We also noted that h was the mean of the uncensored values and the pseudo-observations.

Hence we could change the definition of h to this mean value which keeps invariant the value of the predictive function for given α. However h_j would now be altered to

$$\begin{aligned} h_i' &= (N-1)^{-1}[N\bar{x}_d + (N-d)N\, d^{-1}\, \bar{x}_{N-d} - x_i\,] && \text{for } j = 1, \dots, d \\ &= \bar{x}_d + (N-d)d^{-1}\bar{x}_{N-d} - (N-1)^{-1}x_i && \text{for } j = d+1,\dots,N. \end{aligned} \qquad (3.27)$$

The solution for α^* is now obtained by substituting h_i' for h_i in (3.25).

An unweighted solution in this case is, Geisser (1975b),

$$\hat{\alpha}^* = \frac{N(g-h)^2 - A}{N(g-h)^2 + (N-1)^{-1}A} \text{ for } \hat{\alpha} > 0 \tag{3.28}$$

$$= 0 \quad \text{for } \hat{\alpha} \leq 0$$

where

$$(N-1)A = (d-1)s_d{}^2 + d^{-1}(N-d)^2\,\bar{x}^2_{N-d} + \Sigma^N_{i=d+1}x_i{}^2. \tag{3.29}$$

However, though very simple, this does not appear to be a very satisfactory solution to the problem.

In both (3.24) and (3.27) it is required that $d > 1$ and N-$d > 1$. If $d = 1$ and $N > 2$ then the solution for α^* is the ratio of the second terms in (3.25) utilizing either h_i or h_i' respectively. For $d > 1$, $N = d+1$, the solution is the ratio of the first terms.

4. THE ALTERNATIVE APPROACH-SAMPLE REUSE ALGORITHMS

The second general approach described in Section 1 is both conceptually easier to apply and more readily facilitates arithmetic solutions. We now apply it to the censored situation of the previous section. Using (2.9)

$$\tilde{\alpha} = \frac{t^2(\alpha) - 1}{N} \tag{4.1}$$

where from (3.5)

$$x_i'(\alpha) = x_i + \frac{(d\bar{x}_d + (N-d)\bar{x}_{N-d})\alpha + g}{\alpha\, d+1} \qquad j > d. \tag{4.2}$$

Let

$$\bar{x}(\alpha) = \tfrac{1}{N}[\Sigma^d_{j=1}x_i + \Sigma^N_{i=d+1}x_i'(\alpha)] = \bar{x} + \frac{(N-d)}{N}\left(\frac{N\bar{x}\alpha + g}{\alpha d + 1}\right) \tag{4.3}$$

where $N\bar{x} = \Sigma^N_{\alpha=1}x_i$. Let

$$\beta = \frac{N-d}{N}\left(\frac{N\bar{x}\alpha + g}{\alpha d+1}\right) \tag{4.4}$$

$$(N-1)s^2(\alpha) = \Sigma^d_{j=1}(x_j-\bar{x}-\beta)^2 + \Sigma^N_{j=d+1}(x_j + \beta - \bar{x} - \beta)^2 \tag{4.5}$$

$$= (N-1)s^2 + d\beta^2 - 2\beta d(\bar{x}_d-\bar{x})$$

where $(N-1)s^2 = \Sigma^N_{i=1}(x_j-\bar{x})^2$. Now by definition

$$t^2(\alpha) = \frac{N(\bar{x}(\alpha)- g)^2}{s^2(\alpha)}. \tag{4.6}$$

Hence substituting (4.6) in (4.1) and solving for α in terms of β; i.e.,

$$\mathrm{N}\alpha + 1 = \frac{(\mathrm{N}-d)(g-\bar{x}-\beta)}{d\beta-(\mathrm{N}-d)\bar{x}} \tag{4.7}$$

we obtain a quadratic equation in β

$$a\beta^2 + b\beta + c = 0 \tag{4.8}$$

where

$$\begin{aligned} a &= \mathrm{d}(\mathrm{N}^2-d)/(\mathrm{N}-1) \\ b &= 2(\mathrm{N}-d)d(\bar{x}-\bar{x}_d)(\mathrm{N}-1)^{-1} + d\mathrm{N}(\bar{x}-g)-\mathrm{N}(\mathrm{N}-d)\bar{x} \\ c &= (\mathrm{N}-d)s^2 + \mathrm{N}(\mathrm{N}-d)\bar{x}(g-\bar{x})\,. \end{aligned} \tag{4.9}$$

After obtaining the solution $\hat{\beta}$ we solve for $\hat{\alpha}$ from (4.7) and substituting this in (4.2) we obtain the conditional predictor $x(\hat{\alpha})$ and setting $x_i = 0$ the unconditional predictor.

This approach can also be applied to the case given by equations (2.11) and (2.12), namely $f^* = \alpha^*\bar{x} + (1-\alpha^*)g$ for $0 \leq \alpha^* \leq 1$

$$\alpha^* = (t^2(\alpha^*) - 1)/[t^2(\alpha^*) + (\mathrm{N}-1)^{-1}] \tag{4.10}$$

$$t^2(\alpha^*) = \frac{\mathrm{N}(x(\alpha^*) - g)^2}{s^2(\alpha^*)} \tag{4.11}$$

where the assumed conditional predictor is

$$x_j'(\alpha^*) = x_i + \mathrm{N}d^{-1}\bar{x}\alpha^* + (1-\alpha^*)g \tag{4.12}$$

so that

$$\bar{x}(\alpha^*) = \bar{x} + (\text{N-}d)\text{N}^{-1}[\text{N}d^{-1}\bar{x}\alpha^* + (1\text{-}\alpha^*)g] \qquad (4.13)$$

and

$$(\text{N-}1)s^2(\alpha^*) = (\text{N-}1)s^2 + d\beta^{*2}\text{-}2\beta^* d(\bar{x}_d\text{-}\bar{x})$$

where

$$\text{N}\beta^* = (\text{N-}d)\,(\text{N}d^{-1}\bar{x}\alpha^* + (1\text{-}\alpha^*)g) = (\text{N-}d)\,(zd^{-1}\alpha^* + g) \qquad (4.14)$$

or

$$zd^{-1}(\text{N-}d)\alpha^* = \text{N}\beta^*\text{-}(\text{N-}d)g$$

for $z = \text{N}\bar{x} - dg$.

Hence solutions for α^*, say $\hat{\alpha}^*$, are obtained from the cubic equation

$$(\text{N-}1)(1\text{-}\alpha^*)(d + (\text{N-}d^2)\alpha^*)z^2 = \text{N}d^2(\text{N-}1 + \alpha^*)s^2(\alpha^*). \qquad (4.15)$$

Only one value of the cubic will be appropriate for a fixed $\bar{x}$, s^2 and g. Substitution of the appropriate $\hat{\alpha}^*$ in (4.12) yields the conditional predictor $x(\hat{\alpha}^*)$ and setting $x_i = 0$ yields the unconditional predictor.

We now illustrate this approach with some data obtained from Gnedenko, Belyayev and Solovyev (1969, p. 176). A sample of 100 items are tested and time to failure recorded for each up until 500 time units have elapsed (the actual time unit is not given). It is found that during this period 89 items have survived and the recorded failure times for the other 11 are; 31, 49, 90, 135, 161, 249, 323, 353, 383, 436, 477. The total time on test in undetermined units, is 47,187 (inaccurately given as 47,147 by the authors).

Figure I represents a plot of the predicted value of a future time to failure comparing (4.12), substituting $\hat{\alpha}^*$ for α^*, as a function of g, an apriori guessed value, with (4.2), substituting $\hat{\alpha}$ for α, which derives from the more highly structured predictive approach. The two curves exhibit similar shapes except that the interval for disregarding the data is more than twice as wide for the high structured case and the approach to completely disregarding the guess is far slower.

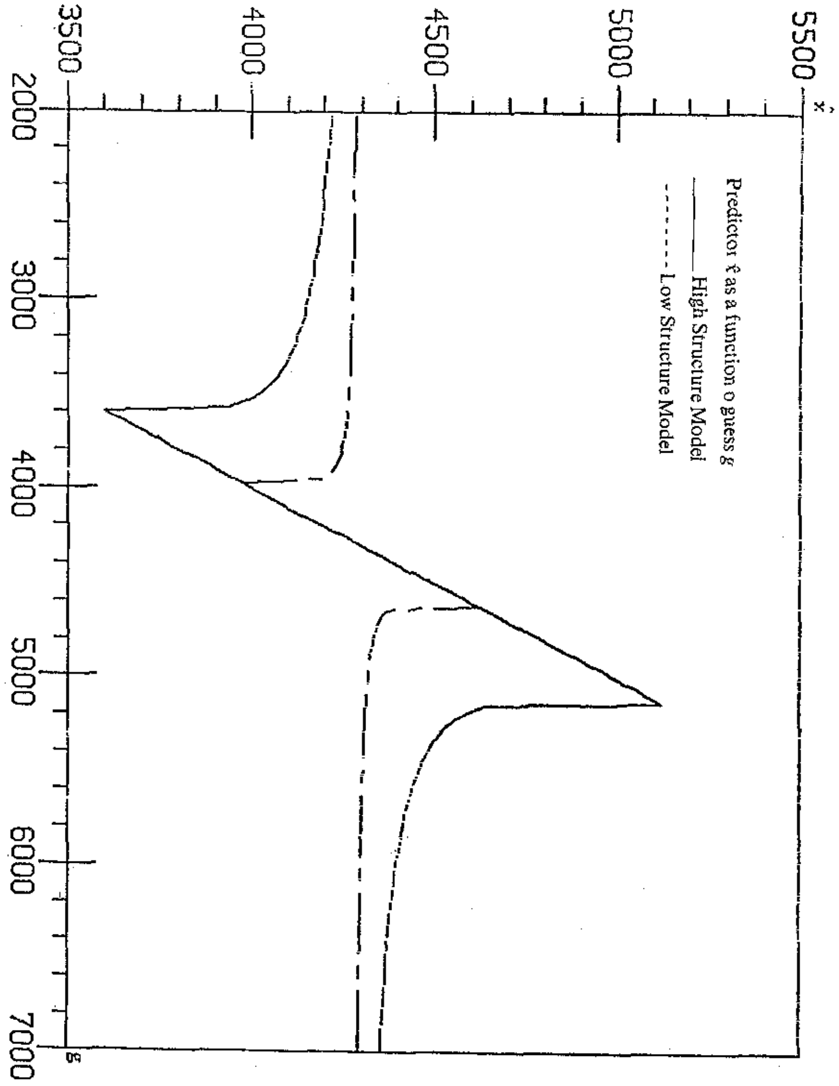


FIGURE 1

Figure II demonstrates how the estimated predictive density of a future observation varies as a function of g using the high structure model. Note that values of g from 3,700 throgh 5,000 result in $\hat{\alpha}=0$ and consequently the density is exponential while for other values of g the density is of the beta form given by (3.2). This accounts for some of the minor perturbations.

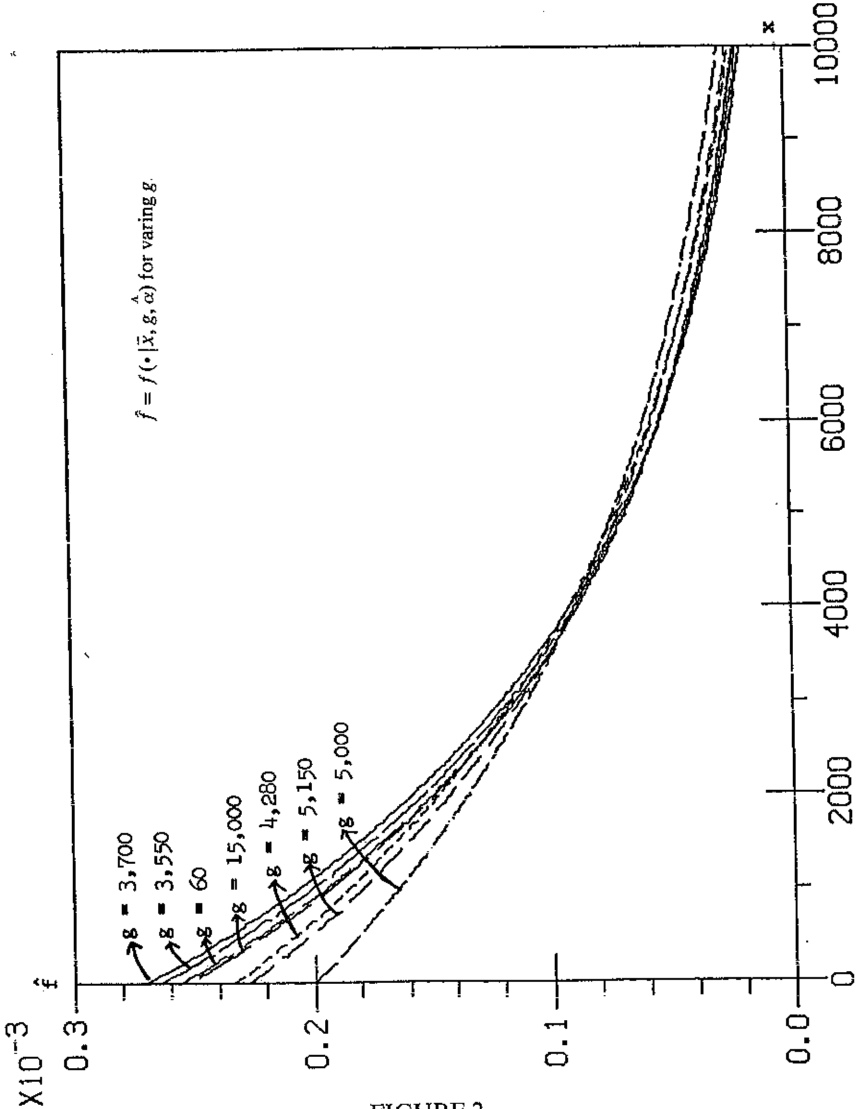


FIGURE 2

Table I gives the shortest .9 probability interval (90th percentile points) for a future value of x for varying g from the estimated predictive distribution.

TABLE I

90th Percentile Point of F(.|x, g, α) to Nearest Integer

g	60	3,550	3,700	4,280	4,290	5,000	5,150	15,000
$\hat{\alpha}$	9.8179	.1259	0	0	0	0	.9409	62.3680
pc point	9,890	9,226	8,520	9,855	9,878	11,513	10,151	10,018

Guesses that are widely discrepant with the data such as 60 and 15,000 are largely ignored and yield percentiles close to that of g = 4290, a guess equivalent to the data predictor. Reversals in percentile points for such values as 3,550 and 3,700 are accounted for by the same phenomenon occuring in Figure 1 and to a lesser extent to the change in the form of the distribution function.

ACKNOWLEDGEMENT

I would like to thank Douglas Kess for the computations and graphs appearing in the previous section. This work was supported by NIH grant GM25271.

REFERENCES

GEISSER, S. (1971) The inferential use of predictive distributions. *Foundations of Statistical Inference.* (B.P. Godambe and D.A. Sprott, eds.) 456-69. Toronto, Montreal: Holt, Rinehard and Winston.

— (1974) A predictive approach to the random effect model. *Biometrika* **61**, 101-107.

— (1975a) The predictive sample reuse method with application. *J. Amer. Statist. Assoc.* **70**, 320-328.

— (1975b) Bayesianism, predictive sample reuse, pseudo-observations, and survival. *Bull. Internat. Statist. Inst.* **40**, 285-289.

— (1975c) A new approach to the fundamental problem of applied statistics. *Sankhya* **37, B** 385-397.

— (1976). Predictivism and sample reuse. *Proc. 21st Design of Experiments Conference* ARO Report 76-2, 385-397.

— (1980) A predictivistic primer. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys.* 363-381 (*A. Zellner, ed*) Amsterdam: North Holland.

GNEDENKO, B.V., BELYAYEV, YU. K. and SOLOVYEV, A.D. (1969) *Mathematical Methods of Reliability Theory*. New York: Academic Press.

STONE, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. B*, **36**, 111-147

DISCUSSION

I. GUTTMAN (*University of Toronto*):

Firstly, I would like to congratulate Professor Geisser for his article and presentation -this is a very stimulating piece of work, and I am honoured to be asked to discuss this paper.

Now I have to report that I have gone through several phases since accepting the invitation to be a discussant of this paper. The first phase said -glance through the paper, get the flavour, report on the flavour, and then, like all good discussants, do the *fungible* thing, talk about my work in the area of reliability, censored data, etc. The second phase, however, intruded, because as I read the paper I became more and more stimulated, interested, frustrated, etc. (underline all of these words) by the underlying ideas. For example, the Sample Re-Use Method is *not* Bayesian and the use of words prior and predictive at various points in the paper are somewhat misleading. (The point that Sample Re-Use is not Bayesian is indeed admitted by Geisser). Indeed, if Geisser had given me permission (how presumptuous can one be) to construct a title, I would perhaps have suggested "Smoothing and Approximations to Properties of Predictive Distributions". Allow me, in the ensuing time to say why.

Suppose the population being sampled has distribution $f(\bullet|\theta)$, and that n independent observations from this population have been taken, say $\mathbf{y}' = (y_1,\ldots,y_n)$. Then, if additionally we are to observe a (future) observation from this population, the conditional distribution h of y, given $\mathbf{y}$, is, using the rules of probability, given by

$$h\,(y|\mathbf{y}) = \int_{\theta\epsilon\Omega} f\,(y|\theta)p\,(\theta|\mathbf{y})d\,\theta \tag{1}$$

where the posterior $p\,(\theta|\mathbf{y})$ is such that

$$p\,(\theta|\mathbf{y}) = c(\mathbf{y})q\,(\theta)\ell(\theta|\mathbf{y}) \tag{2}$$

with

$$\ell(\theta|\mathbf{y}) = \Pi_{i=1}^{n} f\,(y_i|\theta) \text{ and } [c(\mathbf{y})]^{-1} = \int_{\theta\epsilon\Omega} q(\theta)p\,(\theta|\mathbf{y})d\,\theta\,. \tag{3}$$

Here, of course, $q(\theta)$ is the prior distribution of θ and summarizes all the information available to the experimenter *prior* to the taking of the data $\mathbf{y}$. Now suppose the prior q itself depends on certain constants α, i.e.,

$$q(\theta) = q(\theta|\alpha) = q(\theta|\alpha_1,\alpha_2) \tag{4}$$

Then we should write (1) as

$$h(y|\mathbf{y};\alpha) = \int f(y|\theta)p(\theta|\mathbf{y};\alpha)d\theta \tag{5}$$

Now Geisser proceeds as follows: Suppose indeed that the distribution function $q(\theta|\alpha_1,\alpha_2)$ is the "prior" for θ, *and* that α_1 is known, but α_2 is *unknown*, and that $\mathbf{y}$ is observed. Consider the discrepancy function $D(\mathbf{y};\alpha_1;\,\alpha_2)$. Select for α_2 *that value* $\hat{\alpha}_2$ which is such that

$$D(\mathbf{y};\alpha_1,\hat{\alpha}_2) = \min_{\alpha_2} D(\mathbf{y};\,\alpha_1,\,\alpha_2) \tag{6}$$

Note that

$$\hat{\alpha}_2 = \hat{\alpha}_2(\mathbf{y};\alpha_1) \tag{7}$$

(For an example of D, see (2.7) of Section 2 of Geisser's paper). Then, use this to construct a function

$$h^{(s)}(y|\mathbf{y}) = [c(\mathbf{y})]\int f(y|\theta)q(\theta|\alpha_1,\hat{\alpha}_2)\ell(\theta|\mathbf{y})d\theta = h^{(s)}(y|\mathbf{y};\alpha_1;\hat{\alpha}_2) \tag{8}$$

and needless to say

$$h^{(s)}(y|\mathbf{y}) = \int f(y|\theta)p(\theta|\mathbf{y};\alpha_1,\hat{\alpha}_2)d\theta \neq h(y|\mathbf{y}), \tag{8a}$$

where

$$h(y|\mathbf{y}) = \int f(y|\theta)p(\theta|\mathbf{y};\alpha_1,\alpha_2)d\theta \tag{8b}$$

(The superscript (s) in (8a) stands for smoothing $f(y|\theta)$ with $p(\theta|\mathbf{y};\alpha_1,\hat{\alpha}_2)$). Now in (8b), the Bayesian and/or his client is using that value of α_1 and α_2 that arises due to prior information about θ, and in general, this choice will be different than $(\alpha_1,\hat{\alpha}_2)$ — in fact, the $\hat{\alpha}_2$'*s* themselves may vary according to the different nature of the choice of D. This point aside, we are now asked to make the step that regards $h^{(s)}$ as an approximation to h, i.e.,

$$h^{(s)}(y|\mathbf{y}) \approx h(y|\mathbf{y}) \tag{8c}$$

This, it seems to me, must be justified.

Note that we are using a "prior" in (8) which is a function of the data $\mathbf{y}$, a violation of the cannon of Bayesianism which loosely speaking says that if we are to be coherent then the *prior cannot depend on the data*. In fact when we use (8), it seems to me that we can ligitimately ask is there a better way of smoothing $f(y|\theta)$ than by use of $p(\theta|\alpha_1,\hat{\alpha}_2)$? Incidentally, I gather that the use of the term "Predictive Sample Reuse" in the title comes up here because we are using the data not only in the functional form of (2), but also through the use of (6) to use $p(\theta\,|\,\mathbf{y};\,\alpha_1,\hat{\alpha}_2)$ as the smoothing function in $h^{(s)}$, so that even here in the uncomplicated case of no censored data, all observations recorded, we have used the data twice.

But this is compounded in the censored case. In this case the data has a certain structure, viz:

$\mathbf{y}_1' = (y_1, \ldots, y_{n-d})$ are recorded (9)
$\mathbf{y}_2' = (y_{n-d+1}, \ldots, y_n)$ is such that it is known only that $y_j > a_{2j}$ where the a_{2j}, $j = n\text{-}d + 1, \ldots, n$ are known constants.

Using the superscript (c) to denote the presence of censored data as in (9), we have that the predictive distribution of y, given censored data is

$$h^{(c)}(y\,|\,\mathbf{y}_1,\mathbf{a}_2) = \int f(y\,|\,\theta)p(\theta\,|\,\mathbf{y}_1,\mathbf{a}_2;\alpha_1,\alpha_2)d\theta \tag{10}$$

where here the posterior p is given by

$$p(\theta\,|\,\mathbf{y}_1,\mathbf{a}_2;\alpha_1,\alpha_2) = k(y)q(\theta\,|\,\alpha_1,\alpha_2)\ell(\theta\,|\,\mathbf{y}_1;\mathbf{a}_2) \tag{11}$$

with

$$\ell(\theta\,|\,\mathbf{y}_1;\mathbf{a}_2) = \{\Pi_{i=1}^{n-d} f(y_i\,|\,\theta)\}\{\Pi_{j=n-d+1}^{n}\,[1\text{-}F(a_{2j}\,|\,\theta)]\} \tag{12}$$

where F is the cumulative of the distribution $f(.\,|\,\theta)$. Geisser now proceeds as follows. Remove a_{2j} from the likelihood and let $\mathbf{a}_2^{(j)}$ be the (d-1) vector obtained from $\mathbf{a}_2$ by deleting a_{2j} from it. Find $h^{(c)}\,(y\,|\,\mathbf{y}_1,\mathbf{a}_2^{(j)})$ using the prescription (10) etc. From this, find the conditional $h^{(c)}(y\,|\,y > a_{2j};\mathbf{y}_1,\,\mathbf{a}_2^{(j)})$ and obtain

$$\mathrm{E}^{(c)}(y\,|\,y > a_{2j}) = y_j^* \tag{13}$$

the conditional expectation of the predictive variable y, given that $y > a_{2j}$, and where we are given the structure (9) with (d-1) censored observations etc.
The set $(y^*_{n-d+1}, \ldots, y^*_n)' = \mathbf{y}_2^*$ is then used along with $\mathbf{y}_1$ in a discrepancy function D^*, to produce a value α_2^* that is such that

$$D^*(\mathbf{y}_1,\mathbf{y}_2^*;\alpha_1,\alpha_2^*) = \min_{\alpha_2} D^*(\mathbf{y}_1,\mathbf{y}_2^*,\alpha_1,\alpha_2) \tag{14}$$

(An example of D^* is the weighted discrepancy function defined at (3.10) of which I will have something to say below.) The value α_2^* so obtained is then used, as before, to obtain (see (10))

$$h^{(c,s)}(y\,|\,\mathbf{y}_1,\mathbf{y}_2^*;\,\alpha_1,\hat{\alpha}_2^*) = \int f(y\,|\,\theta)p(\theta\,|\,\mathbf{y}_1,\mathbf{y}_2^*;\alpha_1,\hat{\alpha}_2^*)d\theta \tag{15}$$

Note that we are re-using the data ($d + 1 + 1$) times — d times to find the set $\mathbf{y}_2^*$ from (13), one additional time in (14) and of course a further additional time in $p(\theta\,|\,\mathbf{y}_1,\mathbf{y}_2;\alpha_1,\hat{\alpha}_2)$.

To proceed further with this discussion, allow me to start at the beginning of Geisser's paper. We are considering the case of addressing an exponential process, whose distribution is given by

$$f(y\,|\,\sigma) = \sigma^{-1}\exp[-y/\sigma],\ y,\sigma > 0 \tag{16}$$

We are assuming that before the taking of sample information taken from the above (single) exponential, the applicable prior of σ is such that

$$p(\sigma) \propto \sigma^{-(\delta+1)}\exp[-\gamma/\sigma],\ \sigma,\gamma,\delta > 0 \tag{17}$$

This of course implies that, a-priori

$$2\gamma/\sigma = \chi^2_{2\delta} \text{ or } \sigma = 2\gamma/\chi^2_{2\delta} \tag{18}$$

and that the prior expectation and variance are

$$\begin{aligned} &\text{(i) } \mathrm{E}(\sigma) = \gamma/(\delta\text{-}1) = g,\ \delta > 1 \\ &\text{(ii) } \mathrm{V}(\sigma) = g^2/(\delta\text{-}2),\ \delta > 2 \end{aligned} \tag{19}$$

In practice, the prior comes "armed" with fixed values of γ and δ, or g and δ, fixed by prior sample information or "experimenter's expertise", and this does touch on the problem of determining whether (17) is applicable, and if so, how to use prior information or experimenter's expertise to arrive at a suitable choice of (g,δ) or (γ,δ) etc. I do not go into this here, but assume that we do have (17) available and that (γ,δ) represent the values chosen (wisely) by the experimenter.

Note that I use the parametrization given in (16), but of course if we let $\mu = 1/\sigma$, we obtain Geisser's formulation. I prefer using σ as in (16) since $E_f(y\,|\,\sigma) = \sigma$. Note again that $p\,(\sigma)$ exists for $\delta > 0$, $E_p(\sigma)$ exists for $\delta > 1$, and that $V_p(\sigma)$ exists for $\delta > 2$. Also, we make note of the fact that if δ and γ tend to zero such that $\gamma/(\delta\text{-}1)$ tends to g, then $p\,(\sigma)$ tends to $p_{ni}(\sigma)$ which is such that

$$p_{ni}(\sigma) \propto 1/\sigma$$

the so called and much maligned *non*-informative prior (*ni*) for σ.

Now an interesting and somewhat novel theme of the author intrudes at this point, and that is the calculation and fitting of the predictive $h\left(y\,|\,{}^{no}_{data}\right)$, *based* on the prior alone, where

$$h\left(y\,|\,{}^{no}_{data}\right) = \int f\,(y\,|\,\sigma)p\,(\sigma)d\sigma \tag{21}$$

which after using (16) and (17) yields

$$h\left(y \mid {}^{no}_{data}\right) = \delta\gamma^{\delta}(\gamma + y)^{-(\delta+1)}, \tag{21a}$$

that is, a priori, we predict y to behave as a scaled Snedecor-F variable, i.e., *a priori*

$$y = (\gamma/\delta)\ F_{2,2\delta} \tag{21b}$$

Note that the mean and variance of this distribution are

$$\text{(i)}\ E_h(y) = \frac{\gamma}{\delta-1} = g = E_p(\sigma)\,, \qquad \delta > 1 \tag{22}$$

and if $\delta > 2$,

$$\text{(ii)}\ V_h(y) = g^2\ \frac{\delta}{\delta-2} > g^2\ \frac{1}{\delta-2} = V_p(\sigma)\,.$$

Now the moments show that fitting the parameter (γ,δ) or (g,δ) using (21) is associated with a distribution that is located at the same place as the prior $p\ (\sigma)$, but has *larger variance* (by a factor δ ; which could be considerable), and the moments of h are functions of the parameter of the prior. A person who would want to nail down information about prior parameters by fitting his information about (γ,δ) through h (which has larger variance) rather than through the prior, must believe in putting the cart before the horse, and notice too that the experimenter is asked to examine his experience and relate it to future y's based on h, which is not based on current experimental data - I doubt that many experimenters would do this.

Now what is going on can be summarized by the following tableau (We shall let $\delta = (1/\alpha) + 1$ or $\alpha = (\delta\text{-}1)^{-1}$.)

A-Priori

Predictive $y = \frac{\gamma}{\delta}\ F_{2,2_\delta}$

Prior on σ: $\sigma = \dfrac{2\gamma}{\chi^2_{2_\delta}} = \dfrac{\gamma}{\delta}\ \dfrac{2\delta}{\chi^2_{2_\delta}}$

or $\quad \sigma = \dfrac{\gamma}{\delta} \lim\limits_{m\to\infty} \dfrac{\chi^2_m/m}{\chi^2_{2_\delta}/_{2_\delta}}$

or $\quad \sigma = \dfrac{\gamma}{\delta} \lim\limits_{m\to\infty} F_{m,2_\delta}$

$$E_h(y) = \frac{\gamma}{\delta-1} = \gamma\alpha = g \qquad\qquad E_p(\sigma) = \frac{\gamma}{\delta} \lim_{m\to\infty} \frac{2\delta}{2\delta-2} = g$$

$$V_h(y) = g^2\ \frac{\delta}{\delta-2} \qquad\qquad V_p(\sigma) = \frac{\gamma^2}{\delta^2} \lim_{m\to\infty} (F_{m,2_\delta})$$

$$\vdots$$

$$= g^2\ \frac{1+\alpha}{1-\alpha} \qquad\qquad = g^2\ \frac{\alpha}{1-\alpha}$$

Note again that we may write

$$V_p(\sigma) = \frac{\gamma^2}{\delta^2} \lim_{m\to\infty} V\,(F_{m,2_\delta}) \tag{23}$$

as

$$V_p(\sigma) = \frac{\gamma^2}{(\delta\text{-}1)^2}\ \frac{1}{\delta\text{-}2} \lim_{m\to\infty} u(m) \tag{23a}$$

$$= g^2\ \frac{1}{\delta\text{-}2} \lim_{m\to\infty} u(m)$$

where $u(m)$ is such that

$$u(m) = 1 + \frac{2(\delta\text{-}1)}{m}$$

so that

$$\lim u(m) = 1 \tag{23c}$$

Now Geisser's method of fitting using h amounts to saying replace $\lim\limits_{m\to\infty}$

$u(m)$, which equals 1, by $u(2) = \delta$, while of course, the Bayesian, who is using that $p\ (\sigma)$ given by (17), that is, σ is *a-priori*, the scales inverted Chi-Square variable given in (18), is using $u(\infty) = 1$.

Having advocated the fitting of the *no-data* predictive, there is what amounts to some backtracking from this position by Geisser, because he now assumes that $g = E_h(y) \big(= E_p(\sigma)\big)$ is assumed known $\big($i.e. picked by soliciting from the experimenter information, sample or otherwise, about $E_h(y)\big)$ and then, rather than continuing with the fitting of h, chooses $\delta = \frac{1}{\alpha} + 1$ by employing the discrepancy function D or D^* and finally the value of α that minimizes the chosen discrepancy function. Here $D = D(\mathbf{y}; g, \alpha)$ is used it all observations recarded, while $D^* = D^*\ (\mathbf{y}_1; \mathbf{y}_2^*; g, \alpha^*)$ is used if there is censored observations -see the previous discussion here and Geisser's paper, relations (2.7) and (3.10). We again note that doing this amount to choosing a value for a parameter of the prior which depends on the data, a cannon of Bayesianism thus being violated.

Indeed, what would a "Strict Bayesian" do in this problem? (I am indebted to George Barnard for pointing out that the definite article "a" instead of "the" should be used before the words "Strict Bayesian"). We suppose that the process being sampled is as given in (16), that the appropriate prior based on the experimenter's experience and knowledge is given by (17) with δ and γ fixed. Now suppose n units are put on test, and that

(i) n_1 observations, say $y_j^{(c)}$, $j = 1,\ldots,n_1$, unrecorded, but known that lifetimes are less than a_1, that is, $y_i^{(c)} < a_1$, $i = 1,\ldots,a_1$;

(ii) $n\text{-}n_1 - n_2$ observations recorded, say $y_j, j = n_1 + 1,\ldots,n\text{-}n_2$; (24)

(iii) n_2 observations, say $y_i^{(c)}$, $t = n\text{-}n_2 + 1,\ldots,n$, unrecorded, but known that lifetimes are greater than a_2, that is, $y_i^{(c)} > a_2$, $t = n\text{-}n_2 + 1,\ldots,n$

(in our previous discussion, $n_1 = 0$ and $y_i^{(c)} > a_{2_t}$, where $a_{2_t} \equiv a_2$; Geisser's illustrative example involves the case $a_{2_t} \equiv a_2$ and that is why I have decide to look at this case at this point).

From (24) we have that the likelihood is such that

$$\ell\,(\sigma\,|\,\mathbf{y}_2; a_1, a_2) \propto [1\text{-}\exp(\text{-}a_1/\sigma)]^{n_1}$$
$$\times\ \sigma^{-(n-n_1-n_2)} \exp(\text{-}t^{(0)}/\sigma) \times\ (\exp(\text{-}a_2/\sigma)^{n_2} \qquad (25)$$

where $t^{(0)} = \Sigma_{i=1}^{n-n_1-n_2} y_i$ is the sum of the recorded observations. We can thus use all the above ingredients and find

$$p\ (\sigma\,|\,\text{data}) = K\,\Sigma_{j=0}^{n_1} \,{}^{n_1}_{\ j}\ (\text{-}1)^{\sigma^{-(n-n_1-n_2+\delta+1)}} \qquad (26)$$
$$\times\ \exp[\text{-}(t^{(0)} + n_2 a_2 + j a_1 + \gamma)/\sigma]\ .$$

Using (26) the results for ultimate calculation of the predictive distributions are as follows:

I. Uncensored Case: $(n_1 = n_2 = 0)$.

Using the previous definitions we find for this case that:

$$p\ (\sigma\,|\,\mathbf{y}) = \frac{(t+\gamma)^{n+\delta}}{\Gamma(n+\delta)}\ \frac{1}{\sigma^{n+\delta+1}}\ \exp - \frac{t+\gamma}{\sigma} \qquad (27)$$

where $t = t^{(0)} = \Sigma_1^n y_i$, so that, a posteriori,

$$2(t+\gamma)/\sigma = \chi^2_{2(n+\delta)} \qquad (27a)$$

This in turn implies that

$$h(y\,|\,\mathbf{y}) = \int_0^\infty f\,(y\,|\,\sigma) p\ (\sigma\,|\,\mathbf{y}) d\sigma \qquad (28)$$
$$= \frac{n+\delta}{t+\gamma}\left(1 + \frac{y}{t+\gamma}\right)^{-(n+\delta+1)}$$

that is, the predictive distribution is such that

$$y = \frac{t+\gamma}{n+\delta}\ F_{2,\,2(n+\delta)}. \qquad (28a)$$

We find

$$\text{(i)} \qquad E(y\,|\,\mathbf{y}) = \frac{t+\gamma}{n+\delta\text{-}1} = \frac{\alpha n\bar{y} + g}{n\alpha + 1} = f \qquad (28b)$$

$$\text{(ii)} \qquad V(y\,|\,\mathbf{y}) = f^{\,2} \frac{n+\delta}{(n+\delta\text{-}1)^2\,(n+\delta\text{-}2)}$$

and it is to be recalled that Geisser assumes g known and picks α to minimize D given by his (2.7), viz.

$$D\ (\alpha) = n^{-1}\,\Sigma_1^n\,[f_i - y_i]^2 \qquad (28c)$$

where f_i has the same form as f in (28b), but leaves y_i out, that is,

$$f_i = \frac{\alpha(n-1)\bar{y}_i + g}{\alpha(n-1) + 1}$$

and $\bar{y}_i = (n-1)^{-1} \sum_{j \neq i} y_j$.

II. **Case of Censoring on the right only** $(n_1 = 0; n_2 > 0)$

Using previous definitions, for this case we find:

$$p(\sigma | \mathbf{y}_2 ; a_2) = \frac{(t^{(0)}+\gamma+n_2\ a_2)^{n-n_2+\delta}}{\Gamma(n-n_2+\delta)} \frac{\exp -\left(\frac{(t^{(0)}+\gamma+n_2\ a_2)}{\sigma}\right)}{\sigma^{(n-n_2+\sigma+1)}}$$

that is, a posteriori

$$2(t^{(0)} + n_2\ a_2 + \gamma)/\sigma = \chi^2_{2\ (n-n_2+\delta)}\ . \tag{29a}$$

Further, the predictive distribution $h^{(c)}$ is such that

$$y = \frac{(t^{(0)} + n_2\ a_2 + \gamma)}{n-n_2 + \delta}\ F_{2,\ 2(n-n_2+\delta)}. \tag{30}$$

Recall that $\gamma = g(\delta\text{-}1) = g/\alpha$. Note too that (γ,δ) $\big($or (g,δ) or $(g,\alpha)\big)$ is specified at the outset by the experimenter. So a Strict Bayesian who wants to do some predicting in this situation uses (30) which is completely specified. Note too, that using (30) and letting $\alpha \to 0$ implies that

$$y = g\,\chi^2_2\,/2. \tag{31}$$

Hence, in particular, we would estimate the 90th percentile of future y's, say $\tilde{y}_{\cdot 10}$, that is, the point exceeded with probability $\cdot 10$ when using the predictive as

$$\tilde{y}_{\cdot 10} = \begin{cases} \dfrac{t^{(0)} + n_2\,a_2 + (g/\alpha)}{n-n_2+1+(1/\alpha)}\ F_{2,2(n-n_2+1)+\frac{2}{\alpha}\ ;\ \cdot 10} & \text{if } \alpha \neq 0, \\[2ex] g\chi^2_{2,\ \cdot 10}/2 = g\ (2\cdot 3026) & \text{if } \alpha = 0. \end{cases} \tag{32}$$

TABLE I

$\tilde{y}_{\cdot 10}$ (to nearest integer)

		g							
α	δ	60	3550	3700	4280	4290	5000	5150	15,000
0	∞	1138	8174	8520	9855	9878	11,513	11,858	34,539
1/10	11	5273	9124	9290	9929	9940	10,724	10,889	21,759
2/10	6	6888	9420	9528	9949	9956	10,471	10,580	17,724
3/10	$4\frac{1}{3}$	7678	9563	9644	9957	9963	10,346	10,427	15,747
4/10	$3\frac{1}{2}$	8147	9648	9713	9963	9967	10,272	10,367	14,575
5/10	3	8457	9705	9759	9966	9970	10,223	10,277	13,799
6/10	$2\frac{2}{3}$	8678	9745	9791	9968	9971	10,188	10,234	13,247
7/10	$2\frac{3}{7}$	8842	9775	9815	9970	9973	10,162	10,202	12,834
8/10	$2\frac{1}{4}$	8970	9798	9834	9971	9974	10,142	10,178	12,514
9/10	$2\frac{1}{9}$	9072	9817	9849	9972	9974	10,126	10,158	12,259
1	2	9155	9832	9861	9973	9975	10,113	10,142	12,050
Geisser	- - -	9890	9226	8520	9855	9878	11,513	10,151	10,018

MLE of σ to the nearest integer is 4290.

[$F_{m_1, m_2, \beta}$ denotes the point exceeded with probability β when using the Snedecor F with (m_1,m_2) degrees of freedom and $\chi^2_{2;.10}$ is the point exceeded with probability •10 when using the Chi-Square distribution with 2 degrees of freedom and is equal to 4•6052.]

We illustrate the different types of results that emerge using Geisser's data. Consulting Geisser's paper, we find

$$n = 100;\ n_2 = 89;\ n - n_2 = 11;\ a_2 = 500; \tag{33}$$
$$n_2\,a_2 = 89\ (500) = 44{,}500;\ t^{(0)} + n_2\,a_2 = 47{,}187;$$
$$\text{MLE} = 47{,}187/11 = 4290\,.$$

Recall that Geisser uses (32) with α replaced by $\hat{\alpha}$, and $\hat{\alpha}$ is obtained by following the procedure described starting at (12). The results are given in Table I. We are assuming *that $\alpha > 0$ (so that the mean g exists) and we have cut off the table at the line α = 1*, but it could continue indefinitely in principle. As the last line, we have inserted Geisser's results. (At this writing Geisser did not supply the values of $\hat{\alpha}$ found, but of course, it is easy to see that for his entries for the cases g = 3700, 4280, 4290 and 5000 that his $\hat{\alpha} = 0$.)

Note that unlike Geisser's 1 line table, there are no reversals along rows in the main body of the table. Further, the columns to the right of the MLE are decreasing and viceversa. And it is interesting to note again that priors do produce the different results indicated by the Table, different from the 1-line table of Geisser's, which after all could be very different itself depending on the type of D^* function used. Note that if Geisser were to use the weighted function given by his (3.10), then a minor quarrel could be picked. In our notation, the weights used are

(i) $$\left[E_\sigma\{V(y|\sigma)\}\right]^{-1} = \left[E_\sigma\{\sigma^2\}\right]^{-1}$$

for the uncensored variables, where the expectation is taken with respect to the posterior of σ given in (29), and

(ii) $$\left[V(y|y > a_2)\right]^{-1}$$

for the censored variables (recall these are all censored at a_2), where here the variance is taken with respect to the conditional predictive $h^{(c)}(y|y > a_2)$ where the unconditional $h^{(c)}$ is specified in (30). *The minor quarrel is with (i) -in most applications $a_{2j} \equiv a_2$ (in* this example $a_2 = 500$) because of a time constraint, or a gauge calibrated between (0, a_2) only, etc., and realistically then, once we see the data the recorded observations are known to be less than a_2. Hence the recommendation would be, not to use (i), but $[V(y|y < a_2)]^{-1}$.

III. Censored observations on the left and right $(n_1 > 0, n_2 > 0)$.

For this case it is easy to show that we may write the posterior as

$$p\ (\sigma|\text{data}) = K\ \Sigma_{j=0}^{n_1}\binom{n_1}{j}(-1)^j\sigma^{-(n-n_1-n_2+\delta+1)}\ \exp(-v_j/\sigma) \tag{34}$$

where

$$v_j = t^{(0)} + n_2\,a_2 + ja_1 + \gamma\,, \tag{34a}$$

and K is such that

$$K^{-1} = \Sigma_{j=0}^{n_1}(-1)^j\binom{n_1}{j}\ \frac{\Gamma(n-n_1-n_2+\delta)}{v_j^{n-n_1-n_2+\delta}} \tag{34b}$$

Using this we may in turn find that the predictive density is given by

$$h^c(y|\mathbf{y}_1^{(c)};y_2;\mathbf{y}_2^{(c)}) = \Sigma_{j=0}^{n_1}\ c_j q(y;v_j;n-n_1-n_2+\delta) \tag{35}$$

where in general

$$q(y;b;c) = \tfrac{c}{b}\left[1 + \tfrac{y}{b}\right]^{-(c+1)}, b,c > 0 \tag{35a}$$

and where

$$c_j = \binom{n_1}{j}(-1)^j\ v_j^{-(n-n_1-n_2+\delta)}\ /\Sigma_{\ell=1}^{n_1}\binom{n_1}{\ell}(-1)^\ell v_\ell^{-(n-n_1-n_2+\delta)} \tag{35b}$$

To illustrate what happens in this case, we have censored Geisser's data on the right at $a_1 = 60$, so that we are pretending that we have the following sample information:

n_1 = 2 observations less than $a_1 = 60$;
n_2 = 89 observations greater than $a_2 = 500$ (36)
n-n_1-n_2 = 9 observations recorded, and observed to be 90, 135, 161, 249, 323, 353, 833, 436, 477.

Note that $t^{(0)}$ = 2,607. Using the above data in (35), and dealing with the case where α = •5 , that is, δ = 3, we find that the 90th percentile of this distribution for various values of g are as given in Table II.

TABLE II 90^{th} percentile for the predictive density (35) based on the data set (36), for $\alpha = \cdot 5$ and g as tabled

g	60	3,550	3700	4280	4290	5000	5150	15,000
90^{th} percentile	8,414	9,666	9720	9927	9930	10,185	10,239	13,782

A comparison with the $\alpha = \cdot 5$ line of Table I yields the (expected!) fact that the corresponding entries are all less than the corresponding entries of Table I. (A program that tabulates the cumulative of (35) is available from the Department of Statistics, University of Toronto).

Finally, I want to congratulate Geisser again. His paper is very throught provoking and has proven to be very stimulating (to this person at least) and some very subtle issues are raised in this paper. To the data analyst and to those who worry about foundations, this work raises some profound questions to which some clear answers are deserved. In the meantime, the methods proposed by Geisser are of great interest, and he is to be congratulated for the inventive procedures that he has developed.

S.J. PRESS (*University of California, Riverside*):

Professor Geisser has provided us with yet another illustration of the versatility of the predictive sample reuse method that he and Professor Mervyn Stone introduced in different forms, independently, in 1974, in their now well-known and celebrated papers that both appeared in England, in *Biometrika*, and in *JRSS* (B), respectively. Professor Geisser has now shown us how to apply this methodology to the prediction of future observations, when some of the sample data are *censored*. The problem here, of course, that makes this application different from his earlier applications is that not all of the data are inmediately available as candidates for deletion, in the basic discrepancy function, because of the censoring.

As a solution to this inherent difficulty, the author proposes that we introduce pseudo observations, obtained by using the expectation of the predictive distribution of a censored observation given that the observation (that is, the observed failure time) exceeds a preassigned value, namely, the censored value. To obtain this predictive distribution we must introduce substantial structure into the problem. It seems that, we must have a likelihood function, and a *bona fide* prior distribution on the unknown parameters. From this structure we obtain a posterior, and subsequently, a predictive distribution. Taking expectations in the latter yields a "pseudo observation"

The author suggests that when we turn the sample reuse crank, we should utilize the pseudo observations as well as the uncensored ones, and he suggests two methods for doing so. He also suggests that the discrepancy function should be formed as a weighted average of the individual discrepancies obtained by deletion of observations, the weights being assigned according to some specific suggestions.

Finally, Professor Geisser has applied his paradigm to some actual failure time data.

I would like now to make some comments and to raise some questions.

1. My first question concerns the parametric structure imposed on the problem. The recommended approach requires that we make parametric assumptions about both the sample data and about the parameters of the sampling distribution. If, we must impose such structure anyhow, as we would do in a conventional frequentist approach, or in a conventional Bayesian approach, why should we utilize the sample reuse method at all, in this application? To do so, we must introduce some ad hockery regarding the form of our discrepancy function, our predictor function, etc. In other applications, we would presumably be trading off some precision of results, as a result of this ad hockery, in order to gain robustness of prediction with respect to distributional assumptions. In this case, what do we gain?

2. I would like now to question the assignment of weights. Isn't the assignment of weights to the discrepancies quite arbitrary? Certainly the assignment is no less arbitrary that the assumptions made about the form of the discrepancy function, the form of the predictor function, etc. On what basis has the author selected the weights? It seems to me that using precisions as weights is motivated by a normal distribution assumption. But in the case where the data are more likely to be some member of a family of non-normal waiting-time distributions (exponential is what Professor Geisser used as an illustration), why use precision weights?

3. The author combines subjective information with sample information, in a more or less Bayesian way, but violates Bayes' theorem by using sample data to assess the parameters of the prior distribution. It seems to me that there has been ample precedent in the literature for this kind of approach, called empirical Bayes. But this raises the natural question, should we use a moment matching assessment technique or perhaps we should use some other method of getting at the parameters, and then do maximum likelihood estimation of the hyperparameters by maximizing the marginal distribution of the data given the hyperparameters? We would of course need to adopt a likelihood function to do this. Perhaps a smaller risk would be obtained, an important consideration for an empirical Bayesian.

4. My last question involves the underlying parameters of the prior distribution again. A gamma prior is suggested in the paper, for the mean of the sampling distribution of failure time. This is a two parameter prior. But the ensuing analysis really involves only the shape parameter and assumes we know the scale parameter. Perhaps the analysis could be carried out for both parameters simultaneously? Perhaps the mathematics is too intractable.

In conclusion, I would like to thank Professor Geisser for an extremely stimulating and thought provoking paper that clearly extends his earlier research in this area, into new and important fields. But given the methodology we have heard about today, as it relates to *censored* data, it seems to me that there is another problem that could probably be treated in an analogous way - this is the problem of *missing data*. We could generate pseudo observations for the missing data and carry out the analysis in like fashion.

Perhaps Professor Geisser will tell us how to do this in one of his future papers on the subjct.

REPLY TO THE DISCUSSION

S. GEISSER (*University of Minnesota*):

In the introduction to my paper I outlined how sample reuse procedures could be executed in the presence of censored data. Two such procedures were suggested, neither requiring distributional assumptions.

Believing that I. J. Good is essentially correct in his view that most reasonable Bayesian applications are inherently compromises with other methods and also that the predictive sample reuse method can be an attractive empirical Bayes procedure - I offered such an application. It was described first for full data sets and then for incomplete data sets with censoring as a particular application at varying levels of inferential structure running the gamut from low to high.

Even in any real subjective application of Bayesian procedures, there comes a point at some level in the possibly infinite hierarchy of hyperparameters and hyperdistributions where one is no longer willing to continue regressing. Among the several alternatives are: (a) assign precise values to some final set of hyperparameters, (b) introduce a so-called non-informative distribution for them, (c) devise an empirical Bayes procedure for their estimation. Given that certain conditions obtain, coherence is guaranteed for (a), problematic for (b), and inevitably vitiated for (c).

My paper, in part, sets forth a new procedure that can be substituted for others useful in (c) and one which has the robust quality of simulating to a large degree on the available data what it requires from a predictor. Although originally the predictive sample reuse method was introduced to provide point predictors for low structure paradigms, here its effectiveness is amply demonstrated as a useful empirical Bayes estimator of a hyperparameter, an intermediate step towards prediction for a high structure paradigm. In particular, a situation is described where it turns out to be easiest and most convenient to apply amongst the usual estimators of this type. For example, in the uncensored situation, we easily obtain the marginal density of $X_1,\ldots X_N$ to be

$$f(x_1,\ldots,x_N|\delta,\gamma) = \frac{\Gamma(\mathrm{N}+\delta)\gamma^{\delta}}{\Gamma(\delta)[\mathrm{N}\bar{x}+\gamma]^{N+\delta}} \qquad (1)$$

where $\bar{x}$ is the mean of the observations.

Assuming $g = \gamma/(\delta\text{-}1)$ is known and transforming to $Y_i = g^{-1} X_i$, the marginal density of $Y_1, \ldots, Y_N$ is

$$f(y_1,\ldots,y_N|\delta) = \frac{\Gamma(N+\delta)(\delta-1)^{\delta}}{\Gamma(\delta)[N\bar{y}+\delta-1]^{N+\delta}} \qquad (2)$$

Here $S = \Sigma_{i=1}^{N} Y_i$ is sufficient for δ and clearly $(\delta - 1)^{-1}S$ is distributed as $\beta_2(N,\delta)$, a beta distribution of the second kind. The method of moments fails because $E(S) = N$ and using the second moment restricts the range of δ. Hence it would require that $\hat{\delta} > 2$, which results in the estimating method imposing a restriction unassumed by the model. The maximum likelihood estimator is a solution to the unwieldy equation

$$\log \frac{\delta\text{-}1}{s+\delta\text{-}1} + \frac{\delta}{\delta\text{-}1} - \frac{N+\delta}{s+\delta\text{-}1} + \Sigma_{j=0}^{N\text{-}1} \frac{1}{\delta+j} = 0 \qquad (3)$$

the number of whose terms increases with the sample size. So much for competitors in terms of ease in getting a sensible estimator.

Professor Press wonders if the analysis could be carried out for both g and δ unknown. For that case, there is no apparent relief for method of moments and maximum likelihood procedures when applied to (1). The PSR method requires solving a cubic equation in the uncensored case and is somwhat more complicated in the censored case. When a given value for g is specified (which is much more likely to be specifiable than δ), the PSR solution as described in the paper is explicit for the uncensored case and easy to achieve in the censored case using the recursive algorithms of section 4 and has the appealing property of being similar to a "testimator"

It is a rare event indeed when a discussion comes perilously close to exceeding the length of the paper at issue. Even rarer when the discussant begins and ends with the same litany of praise and yet the author must disagree with most of the views expressed. I refer, of course, to Professor Guttman's critique. First an exception: I applaud his use of the term *fungible* which I introduced in an attempt to extend exchangeable. It has indicated that a prediction made when a colleague expressed his aversion to such a singularly unattractive word, may yet take hold. Mustering my most somber demeanor, I portentously responded, "It will grow on you."

First we address some minor details. Professor Guttman's equation (6) is meaningless unless a predictive function is specified. A demonstration of consistency; i.e., that the two sides of (8c) approach the same density as the sample size increases, does not present any difficulty. Although I already responded in part to the great to-do about violating primordial Bayesian canons, still permit me to take this opportunity to expose a further serious transgression on my part. To assume that a prior depends on the likelihood is, of course, original sin itself in this theology. Apparently undetected by Professor Guttman, who usually performs yeoman service as a sort of Bayesian superego, was my use of a conjugate prior density - *mea culpa.* Professor Guttman admonishes me for a prior that comes only "one-armed" instead of what he considers to be appropriate - the investigator determining exact values for both hyperparameters. We obviously describe different situations.

Now to more serious questions. I must take very strong issue with his horse-cart analogy. It derives, I believe, from a fundamental misunderstanding of the practical value of parametric infusions into statistical paradigms. Parameters are basically artifices introduced by the statistician to lubricate the modeling procedure, and of course, hyperparameters even more so. In most instances, they are completely alien to the experimenter's thinking who works with and thinks about observables. Hence, if properly questioned he can respond in those terms directly. If you want to elicit more than just a curious stare, try explaining a hyperparameter to an investigator; it is a sure ticket to non-communication. Further, the exercise on predictive and prior variances which has exercised Professor Guttman invites exorcism. They are irrelevant calculations devoid of purpose and meaning in regard to the issues.

Professor Guttman has taken the trouble to calculate tables of the 90th percentile points of the predictive distribution for varying but known α and g and claims to have uncovered the fatal flaw (certain reversals in the probabilities) in using an empirical Bayes procedure - the fault being that it is not "Bayesian." He could have saved himself the trouble by discerning from the table and graphs in the original paper that this had to be the case. On the one hand, these reversals actually demonstrate the fact that when the guessed value of g is very far from the experimental data, the sample reuse procedures wisely discount the value to a greater and greater extent as if g were the product of a demented prior opinion. On the other hand, when the mean of the sample values is within a certain small interval of g, the procedure behaves as if μ were known to be g^{-1} from the start. This is the "testimator" quality of the procedure - it makes every effort to temper the rigidity of coherence with the facts embodied in the data.

Professor Press complains about my weight functions. If he has a better scheme, I would be happy to entertain it because the plethora I presented complicate the procedure far too much. In fact, the more information used, the greater the computational complexity. Even if a set of weights, indisputably appropriate and yielding a reasonably computable solution, were adduced, which is unlikely, I believe the algorithmic method would still be preferable. This was fully described in section 4 and illustrated for the data set. Hence, I echo his complaint but for different reasons.

On the other hand, Professor Guttmam insists that weights be based on a predictive variance conditioned on the observable being less than a given value when in fact it is known that it exceeds that value. This logical inversion indeed makes even a cart-horse analogy pale by comparison.

Trab. Estadist. (1980)
Vol. **31**, pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 469-520 Valencia, Spain: University Press.

11. Beliefs about beliefs

INVITED PAPERS

DICKEY, J.M. (*University College of Wales, Aberystwyth*)
Beliefs about beliefs, a theory of stochastic assessments of subjective probabilities

GOOD, I.J. (*Virginia Polytechnic and State University*)
Some history of the hierarchical Bayesian methodology

DISCUSSANTS

DEGROOT, M.H. (*Carnegie-Mellon University*)
NOVICK, M.R. (*The University of Iowa*)
GEISSER, S. (*University of Minnesota*)
LEONARD, T. (*University of Warwick*)
LINDLEY, D.V. (*University College London*)

REPLY TO THE DISCUSSION

Beliefs about beliefs, a theory for Stochastic Assessments of Subjective Probabilities

J.M. DICKEY
University College of Wales, Aberystwyth *

SUMMARY

Parameterized families of subjective probability distributions can be used to great advantage to model beliefs of experts, especially when such models include dependence on concomitant variables. In one such model, probabilities of simple events can be expressed in loglinear form. In another, a generalization of the multivariate t distribution has concomitant variables entering linearly through the location vector. Interactive interview methods for assessing this second model and matrix extensions thereof were given in recent joint work of the author with A.P. Dawid, J.B. Kadane and others. In any such verbal assessment method, elicited quantiles must be fitted by subjective probability models. The fitting requires the use of a further probability model for errors of elicitation. This paper gives new theory relating the form of the distribution of elicited probabilities and elicited quantiles to the form of the subjective probability distribution. The first and second order moment structures are developed to permit generalized least squares fits.

Keywords: SUBJECTIVE PROBABILITY; PROBABILITY MODELING; PROBABILITY ASSESSMENT; PROBABILITY ELICITATION; STOCHASTIC ELICITATIONS.

1. SUBJECTIVE PROBABILITY MODELS

Mathematically, subjective probability models resemble the more familiar sampling theory models. The usual Kolmogorov axioms will be satisfied by a probability mass or density function, which is nonnegative, integrates to unity and is otherwise well behaved.[1] The distinguishing

* Present affiliation: State University of New York, Albany

[1] Some writers on subjective probability prefer to work in terms of finitely additive probabilities. This distinction is not material to the present paper.

characteristic of a subjective distribution, then, is not some mathematical property, but rather its use to describe a person's state of mind or subjective uncertainty concerning particular events or quantities of interest. Although a realistic subjective probability distribution for a future sample typically has the mathematical property of exchangeability, or another weakened version of the i.i.d. or related property, my emphasis here is on the distinction according to interpretation or use.

Models in general can be classified as fixed or parametric. A *fixed* probability model is a single probability distribution for a scalar or vector random quantity, or a single distribution-valued function of concomitant variables. In the case of subjective probability, the distribution would be said to be *conditional* on the information in the concomitant variables. A *parametric* model, on the other hand, is a class of models indexed by one or more parameters, whose values serve to specify corresponding fixed models in the class. The difference between a concomitant variable and a parameter, for subjective probability, is that a parameter is used to indicate a class of fixed models merely for mathematical convenience. A parameter may fail to have any interpretation as real information. We shall refer to the parametric and fixed forms of the following subjective probability models.

Model 1. Loglinear odds for an event. A person may be uncertain regarding the occurence of a particular event of interest, say for a dichotomous variable $y = 0, 1$, the event $y = 1$. Conditionally on the vector of concomitant variables **x**, his probability is said to take the loglinear form

$$\begin{aligned} p &= \mathrm{Prob}\{y=1\} \\ &= e^u/(1+e^u), \end{aligned} \tag{1.1}$$

where for $\mathbf{x} = (x_1,\ldots,x_r)'$ and $\mathbf{b} = (b_1,\ldots,b_r)'$,

$$u = \mathbf{x}'\mathbf{b} = x_1b_1 + \ldots + x_rb_r. \tag{1.2}$$

Inversely, $u = \ell n\{p/(1\text{-}p)\}$. The corresponding parametric model has the vector of parameters **b**.

Model 2. Location-scale density for a continuous quantity, with linear location and gathered elliptical symmetry. The person may be uncertain about a particular continuous quantity y. His probability distribution is modeled in location-scale form. Suppose it has a density $p\,(y)$, expressible in terms of some special standardized density f,

$$p\,(y) = f\,(\frac{y\text{-}m}{c})\,/c. \tag{1.3}$$

It is as if there were a standard random quantity z having density f, for which

$$y = m + cz. \tag{1.4}$$

The parameters are m and c.

In the presence of the vector of concomitant variables $\mathbf{x} = (x_1,\ldots,x_r)'$ the conditional distribution of y has the linear-form location, for $\mathbf{b} = (b_1,\ldots,b_r)'$,

$$m = \mathbf{x}'\mathbf{b}. \tag{1.5}$$

Then **b** would become the parameter, instead of m. Of course, c too could depend on **x** (and it will in an important case to be introduced).

This model can be usefully extended in various ways. Writing the concomitant vector as an arbitrary function of more elementary variables **h**, $\mathbf{x} = \mathbf{x}(\mathbf{h})$, one has the notion of a subjective response surface. This complements the theory of objective response surfaces as traditionally used in the optimization of industrial processes. The surface ordinate $m(\mathbf{h}) = \mathbf{x}(\mathbf{h})'\mathbf{b}$ would represent a subjective location for the response y, as opposed to an ideal long-term mean response. The location $m(\mathbf{h})$ can serve as a subjective point prediction, while the scale parameter c expresses the amount of predictive uncertainty.

Opinion concerning samples in time can be modeled by replacing the scalars y, m, z by vectors **y**, **m**, **z**; the scale parameter c becomes a matrix C; and if concomitant variables are present, $\mathbf{x}'$ should be replaced by a matrix X whose row vectors are point values for $\mathbf{x}$,'

$$\mathbf{m} = X\mathbf{b}. \tag{1.6}$$

Equations (1.3) through (1.5) then hold again as written with the given replacements. Equation (1.3) for example becomes

$$p\,(y) = f\,\{C^{-1}(\mathbf{y} - X\mathbf{b})\}\{\det(C)\}^{-1}, \tag{1.7}$$

if we assume the matrix C is nonsingular.

Bruce Hill (1969) and A.P. Dawid (1977, 1978) have investigated the property of spherical symmetry, in which the distribution of **z** is invariant under rotations. If $A\mathbf{z}$ would have the same distribution as **z** for any orthogonal matrix A, then the distribution of $\mathbf{y} = \mathbf{m} + C\mathbf{z}$ depends on the scale matrix C only through the product,

$$W = CC' . \tag{1.8}$$

An example of such a location-scale model is the multivariate Student family

$$\mathbf{y} \sim \text{Student}_d(\mathbf{m}, W), \tag{1.9}$$

where $\mathbf{z} \sim \text{Student}_d(\mathbf{0}, \mathrm{I})$ means that $\mathbf{z}$ can be represented as the product of a standard normal vector and the independent random quantity $(d/\chi^2_d)^{1/2}$. Kadane *et al* (1978) have developed such models for subjective probability modeling.

The multivariate Student distribution (1.9) has the property that the density of $\mathbf{y}$ depends on $\mathbf{y}$ only through the positive definite quadratic form $(\mathbf{y}\text{-}\mathbf{m})' W^{-1}(\mathbf{y}\text{-}\mathbf{m})$, and it strictly decreases in this quadratic form. We shall refer to *any* distribution which has these properties as *gathered and elliptically symmetric*. Much of the work here will apply with full force to a general gathered elliptically symmetric distribution with linear location. The main advantage of such models is that they can be maximized in their coefficients vector by the method of generalized least squares. We write for such a model in analogy to (1.9), for $\mathbf{y} = \mathbf{m} + C\mathbf{z}$,

$$\mathbf{y} \sim F(\mathbf{m}, W), \tag{1.10}$$

where $\mathbf{z}$ has the standard distribution $\mathbf{z} \sim F(\mathbf{0},\mathrm{I})$.

Matrix-variate extensions of such models are also available for opinion about multivariate responses sampled at various concomitant points (Dawid, Dickey and Kadane, 1979).

Subjective probability models, such as the models introduced here, are important for situations where there is not a large amount of proper statistical data available and expert opinions must be used for planning experiments or other decision making. Such models are indispensible when there is little or no proper data. Expert opinion is already used extensively now without formal modeling. The intention is that probability models can bring order into expert-opinion processes. The general scientific method urges observation and experimentation where feasible, and samples can be planned and analyzed using subjective probability. But these models are also useful in situations where statistical methods would not be applied.

Modelling of beliefs has the following types of use:

1. Clarification of belief, during the modeling or assesment process.
2. Communication. More precise expression of opinion.
3. Comparison and possible pooling of experts' opinions.
4. Decision; e.g. coherent decision (criterion of maximum utility).
5. Planning of experiments (e.g. criterion of maximum expected value of sample information).
6. Analysis of experimental or observational data. Updating of opinion by probability conditioning.

Given a joint probability distribution for observed data and some uncertain quantities of interest, such as future data, opinion is coherently updated to account for the observed data by the usual probability conditioning in the joint distribution. For example, in the joint distribution (1.6) for $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2')'$, $p(\mathbf{y}_2|\mathbf{y}_1) = p(\mathbf{y}_1,\mathbf{y}_2)/\int p(\mathbf{y}_1,\mathbf{y}_2)d\mathbf{y}_2$. In the multivariate Student case (1.9),

$$\mathbf{y}_2|\mathbf{y}_1 \sim \text{Student}_{d+r_1}\{A(\mathbf{y}_1), B(\mathbf{y}_1)\}, \tag{1.11}$$

where

$$\begin{aligned} A(\mathbf{y}_1) &= \mathbf{m}_2 + W_{21}W_{11}^{-}(\mathbf{y}_1 - \mathbf{m}_1) \qquad (1.12)\\ B(\mathbf{y}_1) &= (1 + r_1/d)^{-1}\{1 + d^{-1}(\mathbf{y}_1 - m_1)' W_{11}^{-}(\mathbf{y}_1 - \mathbf{m}_1)\}\\ &\quad \cdot (W_{22} - W_{21}W_{11}^{-}W_{12}) \quad ; \end{aligned}$$

We have partitioned $\mathbf{m}$ and W conformably to $\mathbf{y}$; W_{11}^{-} is a generalized inverse; and $r_1 = \text{rank } W_{11}$. (Of course, what is actually meant by this in practice is conditioning on a small positive-probability interval for $\mathbf{y}_1$.)

Note that there has been no need to mention Bayes' theorem. It is only in the special case that $p(\mathbf{y})$ is a mixture of sampling models that Bayes' theorem arises. That is, if $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$ in terms of an i.i.d. sampling model $p(\mathbf{y}|\theta)$ with an unknown parameter θ subject to the prior distribution $p(\theta)$, then

$$p(\mathbf{y}_2|\mathbf{y}_1) = \int p(\mathbf{y}_2|\theta)p(\theta|\mathbf{y}_1)d\theta , \tag{1.13}$$

where the posterior distribution in the integrand is obtained by Bayes' theorem,

$$p\,(\theta\,|\,\mathbf{y}_1) = p\,(\mathbf{y}_1|\theta)p\,(\theta)/\textstyle\int p\,(\mathbf{y}_1|\theta)p\,(\theta)\,d\theta \qquad (1.14)$$

A special case of our multivariate Student model (1.9) can be viewed as a subjective average of the familiar normal-linear-regression sampling models in which

$$\mathbf{y}\,|\,\beta,\sigma \sim \text{Normal}(X\beta;\,\sigma^2\text{I}). \qquad (1.15)$$

If β and σ have the usual conjugate prior distribution,

$$\beta\,|\,\sigma \sim \text{Normal}(\mathbf{b};\,\sigma^2 N^{-1}) \qquad (1.16)$$
$$\sigma^2 \sim s^2(d/\chi^2_d),$$

then the corresponding prior-predictive distribution for $\mathbf{y}$ is just the multivariate Student distribution (1.9) with the special parameter values,

$$\mathbf{m} = X\mathbf{b},\quad W = s^2(X'N^{-1}X + \text{I}). \qquad (1.17)$$

The usual Bayesian updating equations for opinion regarding β and σ (Raiffa and Schlaifer, 1961) lead to the same posterior predictive distribution as (1.11) with (1.17). But, of course, our form (1.11) is much more general.

We state again, for emphasis, that a mixture of sampling models is a special case. In the multivariate Student prevision, a special form of the parameter W is implied (1.17), special in the sense that W is then the sum of a scalar matrix and a matrix of rank fixed relative to the sample size (dimensionality of $\mathbf{y}$).

2. THE PROBLEM OF ASSESSMENT

Just as in any mathematical modeling situation, a person who wishes to model his beliefs by probability is faced with the problem of specifying his model. This can be broken down into the subproblems of determining a parametric model and assessing a fixed model within a given parametric model. We treat the latter type of problem here. In practice, the full specification may proceed by an iteration alternating between tasks of the two types.

We assume that the assessor subjectively specifies *aspects* of the model. Aspects may include: probability values; quantiles; moments; even parameters themselves. Typically, he will overdetermine the model by assessing more aspects than are required to fix the model mathematically. That is, his assessed aspects will be logically contradictory under the model, and some kind of fit must be performed. The extent to which they contradict each other can help indicate the degree of suitability of the given parameterized model.

I should like to emphasize here that subjective probability modeling is like any other type of mathematical modeling, in that diagnostic checks are necessary to see whether the chosen parametric model is adequate for the real situation being modeled. Loglinear odds and gathered elliptically symmetric models are here claimed to be widely useful, but like any parametric model, they cannot be universal. (No model is ever exactly true). The main argument for considering them is that they are tractable and allow a wide variety of opinion structures.

We envisage the assessment process as an aspect-specification and fitting cycle:

1. Specify new aspects
2. Fit model to specified aspects
3. Diagnostic checking
4. Change aspects or change parametric model, and go to 1; or stop.

Interactive computer programs for such a process for models of our second type (1.9), (1.17), are reported in Kadane *et al* (1978) and Dickey and Price (1979). This previous work, however, is informal, in using convenient but arbitrary methods for step 2. The present paper attempts to meet the need for reasonable formal criteria and methods for fitting subjective probability models to specified aspects.

A question of interpretation may be of particular interest at this point. The aspect specifications and the model aimed at are both conceived as subjective entities in the sense of being merely expressions of personal opinions, rather than properties of real-world objects or processes. The reader may appreciate, however, that much of the development here would also apply to situations where an underlying probability model, which a person is trying to assess, is considered to have its own objective existence (say, the long term frequency of failure for a particular type of component in an operating nuclear power plant). Then the aspect specifications could be conceived as subjective *estimates* of the objective aspects.

Contexts of the latter sort resemble in many ways the traditional sampling context in which *both* the model and the aspect specifications are objective. That is, data *drawn from* the model are used to form statistics, which then estimate aspects of the model. This resemblance will receive further discussion latter. For the present we merely point out the logical distinction between data *concerning* a model and data *drawn from* a model. The former concept is the more general.

3. STOCHASTIC ASSESSMENT MODELS

We postulate two models, in general. First, the *belief-model* or subjective probability model, denoted p, say a probability mass or density function $p\ (y)$ for the uncertain quantity y. This is the underlying true fixed model, the object of the assessment. It is *true* in the sense of exactly describing the given expert's personal belief, and it takes the form of a probability distribution, possibly conditional on concomitant variables. Aspects, functions of this model, are denoted,

$$u_1, u_2, \ldots, u_n. \tag{3.1}$$

Denote the vector $\mathbf{u} = (u_1, \ldots, u_n)'$. Then $\mathbf{u} = \mathbf{u}(p)$.

Strictly speaking, for the aspects to be functions, the model p would need to be seen as a member of a class of models, such as the class of all distributions for y on the given range. For another example, if the model is parameterized by $\mathbf{a}$, then $\mathbf{u} = \mathbf{u}(\mathbf{a})$. Typically, this function is invertible on a subrange of $\mathbf{u}$ values. For these values, then, the model p would be identified (in the mathematical sense) by $\mathbf{u}$.

The expert assesses values for the aspects,

$$u_1^*, u_2^*, \ldots, u_n^*. \tag{3.2}$$

In vector form, write $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)'$. The second category of model is the *assessment model*, denoted q. This is a probability mass or density function $q\ (\mathbf{u}^*)$ for the random assessments $\mathbf{u}^*$, which depends on the true model p. Whereas $u_1, \ldots, u_n$ "concern" p, $u_1^*, \ldots, u_n^*$ are "drawn from" q. We assume that the dependence of q on p comes only through $\mathbf{u}$, and hence write for given p,

$$q(\mathbf{u}^*) = g\ (\mathbf{u}^*; \mathbf{u}). \tag{3.3}$$

This is a new use for the concept of probability. (See, however, Lindley, Tversky and Brown, 1979). On the one hand, q models the subjective belief of the expert concerning his own belief p. On the other hand, a sample $\mathbf{u}^*$ drawn from q is actually available for analysis, and $\mathbf{u}^*$ can be analysed in any of the ways a statistician would ordinarily work with data drawn from a distribution. The assessment probability for $\mathbf{u}^*$ (3.3) depends on $\mathbf{u}$, and hence on p. Thus, in the case of a parameterized belief model p, the assessment likelihood for the belief parameter $\mathbf{a}$ can be written

$$\ell_q(\mathbf{a}) = q\ (\mathbf{u}^*)_{u=u(a)} = g\ (\mathbf{u}^*; \mathbf{u}(\mathbf{a})\,). \tag{3.4}$$

Consequently, familiar Bayesian or likelihood methods can now be used to make inference concerning p through $\mathbf{a}$. In particular, one can estimate the belief model p by maximizing the assessment likelihood $\ell_q(\mathbf{a})$ (3.4). (The frequentist justifications for maximum likelihood are well known; Bayesians might justify it as an approximate posterior mode).

Lindley, Tversky and Brown (1979) postulate a further probability model in order to carry out Bayesian inference concerning p. For them p itself would be random under a further "prior" distribution.

Example. Assessment likelihood having linear location and gathered elliptical symmetry. Consider the following useful structures for q and p, respectively, in terms of a standard gathered elliptically symmetric distribution $G\ (0,I)$,

1. $\mathbf{u}^* | \mathbf{u} \sim_q G\ (\mathbf{u}, V)$
2. $\mathbf{u} = L\mathbf{a}$.

The assessment likelihood in this case would be maximized by the generalized-least-squares estimate,

$$\hat{\mathbf{a}} = (L'V^{-1}L)^{-1}(L'V^{-1}\,\mathbf{u}^*). \tag{3.5}$$

One usually sees the estimate (3.5) justified by the Gauss-Markov theorem in terms of variance and bias. It was derived here by maximum likelihood. This structure would include the usual normal linear model, to which both such "justifications" apply. Variance, bias, and other moments may fail to exist, however, for more general G.

Note that in the present example very little has yet been stated concerning the belief model p; merely, that some aspects of p are linearly related to some parameters in p. Nothing yet has been assumed regarding the interpretation of $\mathbf{u}$ or $\mathbf{a}$. In a special case of some interest, the object $\mathbf{y}$ of the belief would follow a related subjective-probability model,

3. $\mathbf{y} | \mathbf{u} \sim_p F\ (\mathbf{u}, W)$,

where, for example, $F\,(\mathbf{0,I})$ is the same standard distribution as $G\,(\mathbf{0,I})$. We shall discuss later a possible relevance for taking the matrices V and W to be proportional.

We turn in the following sections to theoretical considerations relating assessment models to the belief models previously given. Particular location and scale structures will be motivated for assessment models q, for use of the generalized least squares estimate (3.5).

4. ASSESSING THE PROBABILITY OF AN EVENT

For aspects, consider the linear logodds of equation (1.1), $u_i = \ell n\{p_i/(1-p_i)\} = \mathbf{x}_i'\mathbf{b}$, $i = 1,\ldots,n$. The expert could assess either u_i or p_i, but we retain the notation in which the logodds are treated as the aspects. In practice, one might prefer to assess p_i directly and then transform to an assessment of u_i. Expanding both the transformation and its inverse about the point $p = \frac{1}{2}$ yields

$$\begin{aligned} u &= 2\{p-(1-p)\} + \tfrac{2}{3}\{p-(1-p)\}^3 + \ldots \\ p &= \tfrac{1}{2} + \tfrac{1}{4}u - \tfrac{1}{48}u^3 + \ldots \end{aligned} \qquad (4.1)$$

Both second-order terms vanish, and so the transformation is approximately linear for moderate probabilities.

Assuming that assessments u^*, p^* are related similarly to u, p, that is by $u^* = \ell n\{p^*/(1-p^*)\}$, we have that unbiasedness of p^* is approximately equivalent to unbiasedness of u^*. Hence we assume for the first moment of u^*.

Assumption 4.1

$$Eu^* = u. \qquad (4.2)$$

Cox (1958) uses the somewhat weaker assumption of a constant bias for u_i^*in the context of subjective estimation of objective probabilities.

We discuss the second moment at length.

It is clear that very small or very large probabilities are assessed with smaller absolute errors than moderate probabilities. We shall argue here for the proportionality

$$\mathrm{Var}(p^*) \propto p\,(1-p) \qquad (4.3)$$

Justification (a). A constant coefficient of variation S.D. $(p^*)/E(p^*)$ would express the idea that the errors in assessment are proportional, in their standard deviation, to the true value p. This seems more reasonable than a constant standard deviation for small probabilities, but perhaps overly optimistic in that such probabilities are notoriously difficult to assess. It would also be unrealistic for large probabilities in not having the standard deviation there smaller than at moderate probabilities. A reasonable compromise which meets all of the above points is to consider the new ratio,

$$\mathrm{S.D.}(p^*)/\{Ep^*)(1-Ep^*)\}^{1/2}. \qquad (4.4)$$

This will be constant under (4.3) for unbiased p^*.

Justification (b). If p^* is Beta distributed under the assessment model, then $\mathrm{Var}(p^*) \propto (Ep^*)(1-Ep^*)$, which again yields (4.3) in the unbiased case.

Justification (c). The variance within the subjective-probability model is $\mathrm{Var}(y) = p\,(1-p)$. We shall argue, below, for the case of continuous y, that assessment variance is proportional to belief variance. By (mere) analogy here, $\mathrm{Var}\,(p^*) \propto \mathrm{Var}(y) = p\,(1-p)$.

Considering now the second moment of u^*, we have, to first order,

$$u^* - u = (du/dp) \cdot (p^* - p). \qquad (4.5)$$

Hence, $\mathrm{Var}(u^*) \doteq (du/dp)^2\,\mathrm{Var}(p^*) \propto \{p\,(1-p)\}^{-2}\{p\,(1-p)\} \propto \{p\,(1-p)\}^{-1}$, by (4.3). This motivates the following:

Assumption 4.2. $\mathrm{Var}(u^*)$ is proportional to $\{p\,(1-p)\}^{-1} = e^{-u} + 2 + e^{u}$.

To use the moment structure of Assumptions 4.1 and 4.2 to fit a loglinear odds model to assessed aspects will require iteration, because the variance is a function of the mean. Further assumptions would also be needed regarding the covariances.

5. ASSESSING QUANTILES OF A LOCATION-SCALE MODEL

For a continuous random quantity y define the πth quantile $(0 < \pi < 1)$ as the number y_π satisfying

$$P\,\{y \leq y_\pi\} = \pi. \qquad (5.1)$$

In the problem of assessing a simple location-scale model (1.3) consider as aspects u_i, the quantiles y_{π_i} for given probabilitiy values π_i, $i = 1,\ldots,n$. A linear relation holds between the quantiles of y and the corresponding quantiles of the standard random quantity z,

$$y_{\pi_i} = m + cz_{\pi_i} \tag{5.2}$$

Hence given the assessed quantiles $y^*_{\pi_i}$, $i = 1,\ldots,n$, a natural method to use to estimate m and c is to fit the straight line (5.2) to the "data", z_{π_i}, $y^*_{\pi_i}$, $i = 1,\ldots,n$. This was proposed by I.J. Good (1978) as a method of reconciling subjective quantiles from several experts in the normal case.

An appealing fitting method to use here is generalized least squares (3.5), and a candidate for the required error-covariance structure will be developed below. Garthwaite and Dickey (1979) study properties of the "bisection" method, a special case of this method when bisection is used for assessing location-scale parameters. In the bisection method, particular quantiles are elicited as medians of distributions conditioned on subintervals.

In the more general multivariate location-scale model with linear form location, $\mathbf{m} = X\mathbf{b}$ (1.7), it might seem reasonable to fit this form for $\mathbf{b}$ after assessing a single quantile of y_i at each point $\mathbf{x}_i$ (y conditional on $\mathbf{x} = \mathbf{x}_i$), where the row vectors $\mathbf{x}_i'$, $i = 1,\ldots,n$, comprise the matrix X. These quantiles $y_{\pi_i}(\mathbf{x}_i)$ would all be assessed for the same probability value say $\pi_i \equiv ½$, an appealing value to use in the elliptically symmetric model (1.9), for which the coordinates of $\mathbf{m}$ are the medians of the coordinates of $\mathbf{y}$. One would fit the linear relation, in this case,

$$y_{.50}(\mathbf{x}_i) = \mathbf{x}_i'\mathbf{b} \tag{5.3}$$

to the "data", $\mathbf{x}_i$, $y^*_{.50}(\mathbf{x}_i)$, $i = 1,\ldots,n$. Again, generalized least squares will require an error-covariance structure.

5.1 *Sample quantiles as estimates of quantiles*

Subjective assessment of quantiles may be preferrable to the assessment of probabilities of intervals or half-lines, because an expert may find it more meaningful to weigh against each other quantities having the same units as the unknown y, relative to a fixed probability, rather than comparing candidate probability values. But how accurate are such quantile assessments? Perhaps a clue is available from the analogous problem of estimating the quantiles of a traditional population by the quantiles of a sample drawn from the population. There is, of course, no *necessary* connection between this and our problem of assessing subjective probability quantiles.

Denote a population by p, or $p\ (y)$. Denote its πth quantile by y_π, and the corresponding quantile of an independent sample from p by y^*_π. Then for large samples, the asymptotic distribution of y^*_π is normal with mean and variance,

$$E(y^*_\pi) = y_\pi$$
$$\mathrm{Var}(y^*_\pi) = \nu^{-1}\pi(1-\pi)/p\ (y_\pi)^2, \tag{5.4}$$

where ν denotes the sample size. Indeed, the joint distribution of the sample quantiles $y^*_{\pi_i}$ at several probability values π_i is asymptotically multivariate normal with the covariance structure,

$$\mathrm{Cov}(y^*_{\pi_1}, y^*_{\pi_2}) = \nu^{-1}\pi_1(1-\pi_2)/\{p\ (y_{\pi_1})p\ (y_{\pi_2})\} \tag{5.5}$$

for $\pi_1 \leq \pi_2$ (Mosteller, 1946).

So sample quantiles are asymptotically unbiased; and in the case of a location-scale family $p(y) = f\{(y-m)/c\}/c$, the variances and covariances will be proportional to the squared population scale parameter,

$$\mathrm{Cov}(y^*_{\pi_1}, y^*_{\pi_2}) = [\nu^{-1}\pi_1(1-\pi_1)/\{f(z_{\pi_1})f(z_{\pi_2})\}]c^2 \tag{5.6}$$

That is, if the distribution being estimated has a variance, the sample quantiles will be distributed with an asymptotic variance proportional to it,

$$\mathrm{Var}(y^*_\pi) \propto \mathrm{Var}(y) \tag{5.7}$$

We shall argue for an analog of this principal in the next section.

In unpublished work Michael Cain has derived assessment fitting procedured for the linear model (1.9), (1.17) using the moment structure of sample quantiles, following a suggestion by J.B. Kadane.

5.2 *Assessed quantiles*

Returning to the general notion of assessed quantiles y^*_π having a distribution q conditional on the quantiles y_π of the distribution p of y, define the cumulative distribution function P for y, at any value y°,

$$P(y^\circ) = \mathrm{Prob}\{y \leq y^\circ\} = \int_{-\infty}^{y^\circ} p(y)dy. \tag{5.8}$$

Then, of course, $\pi = P(y_\pi)$.

Transforming the assessment, define the quantity,

$$\pi^* = P(y^*_\pi) = \int_{-\infty}^{y^*_\pi} p(y)dy. \tag{5.9}$$

In practice, π^* will not be available in numerical form, depending as it does on the model p. But still, π^* is a mathematically well defined random quantity and has a distribution induced by the assessment distribution q, and we can discuss the behavior of π^* relative to the "true" value π.

A main idea of this paper is that the induced distribution of π^* promises to be insensitive to the model p; at any rate, less sensitive than the distribution of the assessed quantile y^*_π itself. The quantity π^* represents the amount of "true" probability included to the left of y^*_π. The assessment errors in y^*_π could be expected to be large if the integrand $p(y)$ of (5.9) is small in the vicinity of y_π, and small if $p(y)$ is large there. The less believable a region is, the more difficult it is to assess a quantile within it, and visa-versa. This would have the effect of stabilizing the distribution of π^* in its dependence on the local behavior of $p(y)$. We consider small errors in y^*_π, and hence small errors in π^*.

Assumption 5.1. π^* is unbiased:

$$E(\pi^*) = \pi. \tag{5.10}$$

Now, take the linear expansion of the cumulative,

$$\pi^* \doteq \pi + p(y_\pi)(y^*_\pi - y_\pi) \tag{5.11}$$

which yields, together with Assumption 5.1.

Consequence 5.2. For small assessment errors, y^*_π is unbiased:

$$E(y^*_\pi) \doteq y_\pi. \tag{5.12}$$

Assumption 5.3. The model $p(y)$ is parameterized as a location-scale family, $y = m + cz$, where z has a known distribution with density $f(z)$ and unit variance. Hence, $p(y_\pi) = f(z_\pi)/\{\mathrm{Var}(y)\}^{1/2}$. (If one makes the assumption that the assessment model $q(y^*_\pi)$ is also of location-scale form, then no moments need exist, and one can read locations and squared-scale parameters for the means and variances throughout this section.)

In spirit similar to Assumption 5.1, we have,

Assumption 5.4. Var (π^*) is constant in m and c.

Consequence 5.5. Proportionality of variances (scales).

$$\mathrm{Var}(y^*_\pi) \doteq \mathrm{Var}(\pi^*)/p(y_\pi)^2 = \{\mathrm{Var}(\pi^*)/f(z_\pi)^2\} \cdot \mathrm{Var}(y) \tag{5.13}$$

Under an assumption analogous to the constant modified coefficient of variation in the linear log odds problem (4.4), we obtain a more explicit form for the dependence on the quantile probability value π, as follows.

Assumption 5.6. Constant moments ratio

$$\mathrm{S.D.}\,(\pi^*)/\{(E\pi^*)(1-E\pi^*)\}^{1/2} = K \tag{5.14}$$

Consequence 5.7.

$$\mathrm{Var}(y^*_\pi) \doteq K^2\pi(1-\pi)/p(y_\pi)^2 = \{K^2\pi(1-\pi)/f(z_\pi)^2\} \cdot \mathrm{Var}\,(y). \tag{5.15}$$

This exhibits an even closer resemblance than (5.13) to the variance of a sample quantile (5.4). It is tempting here to speculate that the covariance of assessed quantiles might have an analogous resemblance to the covariance of sample quantiles,

$$\mathrm{Cov}(y^*_{\pi_1}, y^*_{\pi_2}) = \{K^2\pi_1(1-\pi_2)/f(z_{\pi_1})f(z_{\pi_2})\}\mathrm{Var}(y) \tag{5.16}$$

for $\pi_1 \leq \pi_2$. The corresponding correlation coefficient would be the same as in the sample quantile case, namely $[\{\pi_1/(1-\pi_1)\}/\{\pi_2/(1-\pi_2)\}]^{1/2}$
This correlation approaches unity as $\pi_2 - \pi_1 \to 0$, and so our stochastic assessment model is "smooth" in the assessment of neighboring quantiles. (For sample quantiles, of course, such a limiting operation makes no sense).

We turn finally to the distribution of median assessments $y^*_{50}(\mathbf{x}_i)$ in the gathered elliptically symmetric location-scale model with linear location $y_{50}(\mathbf{x}_i) = \mathbf{x}_i'\mathbf{b}$, $i = 1,\ldots,n$. The essential property for the discussion here is that a set of jointly distributed assessed quantities $y^*_{50}(\mathbf{x}_i)$ have variances proportional to the corresponding jointly distributed observables y_i at $\mathbf{x}_i$. Denote the vectors having these two sets of coordinates, respectively, by $\mathbf{y}^*_{50}$ and $\mathbf{y}$. We extend this property in the following,

Assumption 5.8. In the coordinate system of the principal components of $\mathbf{y}$, the vectors $\mathbf{y}^*_{50}$ and $\mathbf{y}$ again have coordinates with proportional variances: Write $\eta = \mathbf{A}\mathbf{y}$ and $\zeta = \mathbf{A}\mathbf{y}^*_{50}$. Assume that for some orthogonal matrix A, both $\mathrm{Var}(\eta) = \mathrm{Diag}(\tau^2_1, \ldots, \tau^2_n)$ and $\mathrm{Var}\,(\zeta_i) = k\tau^2_i$, $i = 1, \ldots, n$.

Assumption 5.9. Quantities uncorrelated in the belief model correspond to quantities uncorrelated in the assessment model: Assume $\mathrm{Cov}(\zeta_i, \zeta_j) = 0$, $i \neq j$.

Clearly then, the matrices $\mathrm{Var}(\eta) = k\,\mathrm{Var}(\zeta)$, and hence,

Consequence 5.10. Proportionality of covariance matrices:

$$\mathrm{Var}(\mathbf{y}^*_{50}) = k\,\mathrm{Var}(\mathbf{y}) \tag{5.17}$$

To use this moments structure for a fit on the linear location will require, of course, a separate assessment and fitting procedure for the scale matrix, or an iteration alternating between the scale and the location.

6. DISCUSSION

We have argued theoretically for particular forms of probability model for the behavior of subjectively assessed aspects of a probability model of belief or frequency. Such a model of assessment behavior would be essentially *descriptive* in its interpretation, rather than *normative* as the underlying belief model. As such, its suitability should be investigated experimentally. Do errors in assessing belief behave as advertized; or is another stochastic model more realistic; or can better descriptions be given in deterministic form?

One difficulty to be met in the experimental study of assessment models is that of establishing the underlying belief model. Assessments are measurements on beliefs, and to study the distribution of assessment errors would seem to require working in controlled conditions where the "true" opinion values are known, that is, known to the experimenter but not known precisely to the person whose opinions are being assessed. This seems hardly likely for underlying *subjective* probabilities. The subjective estimation of *objective* probabilities is another story, and perhaps experiments on this problem can be extrapolated in their implications to the former problem. Of course, sophisticated statistical methods are also available for inferring the distribution of errors without knowing the underlying "true" values, though typically, this will require additional structural assumptions.

A more fundamental difficulty must, however, be addressed here, and that is that underlying belief models may fail to exist in any realistic sense. In his second philosophy, following Ramsey, Wittgenstein (1953) dealt devastingly with all kinds of logical constructs invented to describe the human mind. A mind's "perceptions" of its own "mental states" (including beliefs) was a favorite target of his. Such logical constructs seem to exhibit what DeFinetti (1974, p.22) calls "the inveterate tendency of savages to objectivize and mythologize everything; a tendency that, unfortunately, has been, and is, favoured by many more philosophers than have struggled to free us from it"

My purpose in this paper is to investigate a framework that may be of use in practice, in the sense that the subjective probability models eventually fixed by an assessment-and-fitting cycle will be found useful. The suspicion remains that the model produced many depend strongly on the assessment method (Hogarth 1975). A person's opinions are not coherent (probabilistic) to begin with, but only as he makes deliberate use of the normative theory of subjective probability. Stochastic assessment models may help provide ways of using the normative theory.

ACKNOWLEDGEMENT

Research supported by a grant from the Social Science Research Council, United Kingdom.

REFERENCES

COX, D.R. (1958). Two further appplications of a model for binary regression. *Biometrika* **45**, 562-65.

DAWID, A.P. (1978). Extendability of spherical matrix distributions. *J. of Multivariate Analysis*, **8**, 559-566.

— (1977). Spherical matrix distributions and a multivarite model. *J. Roy. Statist. Soc., B.* **39** 254-261.

DAWID, A.P., DICKEY, J.M. and KADANE, J.B. (1979). Matrix t and multivariate t assessment. *Tech. Rep.*, Department of Statistics, University College of Wales. Aberystwyth.

DE FINETTI, B. (1974). *Theory of Probability, Vol. 1.* New York: Wiley.

DICKEY, J.M. and PRICE, D.E.(1979). Interactive methods for choice of linear subjective probability models. *Tech. Rep.* Statistics Department, University College of Wales, Aberystwyth.

GARTHWAITE, P.H. and DICKEY, J.M. (1979). Assessment by bisection methods for a location--scale family. *Tech. Rep.* Department of Statistics, University College of Wales. Aberystwyth.

GOOD, I.J. (1978). On the combination of judgements concerning quantiles of a distribution with potential application to the estimation of mineral resources. *Tech. Rep.* Department of Statistics, Virginia Polytechnic Institute and State University.

HILL, B.M. (1969). Foundations for the theory of least squares. *J. Roy. Statist. Soc. B.* **31**, 89-97.

HOGARTH, R.M. (1975). Cognitive processes and the assessment of subjective probability distributions. *J. Amer. Statist. Assoc.*, **70**, 271-94.

KADANE, J.B., DICKEY, J.M., WINKLER, R.L., SMITH, W.S. and PETERS, S.C. (1978). Interactive elicitation of opinion for a normal linear model. *Tech. Rep.* **150**, Department of Statistics, Carnegie-Mellon University. *J. Amer. Statist. Assoc.* Dec. issue, 1980.

LINDLEY, D.V., TVERSKY, A., BROWN, R.V. (1979). On the reconciliation of probability assessments. *J. Roy. Statist. Soc. A.* **142**, 146-180 (with discussion).

MOSTELLER, F. (1946). *On some useful "inefficient" statistics. Ann. Math. Statist.* **17**, *377-408.*

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory.* Boston: Graduate School of Business, Harvard University.

WITTGENSTEIN, L. (1953). *Philosophical Investigations.* Oxford: Basil Blackwell.

Some History of the Hierarchical Bayesian Methodology

I.J. GOOD
Virginia Polytechnic Institute and State University

SUMMARY

A standard technique in subjective "Bayesian" methodology is for a subject ("you") to make judgements of the probabilities that a physical probability lies in various intervals. In the hierarchical Bayesian technique you make probability judgements (of a higher type, order, level, or stage) concerning the judgements of lower type. The paper will outline *some* of the history of this hierarchical technique with emphasis on the contributions by I.J. Good because I have read every word written by him.

Keywords: HIERARCHICAL BAYES; PARTIALLY-ORDERED PROBABILITIES; UPPER AND LOWER PROBABILITIES; EMPIRICAL BAYES; SPECIES FREQUENCIES; MULTINOMIAL ESTIMATION; PROBABILITY ESTIMATION IN CONTINGENCY TABLES; PROBABILITY DENSITY ESTIMATION; MAXIMUM ENTROPY; ML/E METHOD; TYPE II LIKELIHOOD RATIO; INFORMATION IN MARGINAL TOTALS; KINDS OF PROBABILITY; BAYES/NON-BAYES SYNTHESIS; HYPER-RAZOR OF DUNS AND OCKHAM.

1. PHILOSOPHY

In 1947, when few statisticians supported a Bayesian position, I had a non-monetary bet with M.S. Bartlett that the predominant philosophy of statistics a century ahead would be Bayesian. A third of a century has now elapsed and the trend supports me, but I would now modify my forecast. I think the predominant philosophy will be a Bayes/non-Bayes synthesis or compromise, and that the Bayesian part will be mostly hierarchical. But before discussing hierarchical methods, let me "prove" that my philosophy of a Bayes/non-Bayes compromise or synthesis is necessary for human reasoning, leaving aside the arguments for the specific axioms.

Proof. Aristotelean logic is insufficient for reasoning in most circumstances, and probabilities must be imcorporated. You are therefore forced to make probability judgements. These subjective probabilities are more directly involved in your thinking than are physical probabilities. This would be even more obvious if you were an android (and you cannot prove you are not). Thus subjective probabilities are required for reasoning. The probabilities cannot be sharp, in general. For it would be only a joke if you were to say that the probability of rain tomorrow (however sharply defined) is 0.3057876289. Therefore a theory of partially ordered subjective probabilities is a necessary ingredient of rationality. Such a theory is "a compromise between Bayesian and non-Bayesian ideas. For if a probability is judged merely to lie between 0 and 1, this is equivalent to making no judgment about it at all" (Good, 1976b, p.137). Therefore a Bayes/non-Bayes compromise or synthesis is an essential ingredient of a theory of rationality. *Quod erat demonstrandum*.

The notion of a hierarchy of different types, orders, levels, or stages of probability is natural (i) in a theory of physical (material) probabilities, (ii) in a theory of subjective (personal) probabilities, and (iii) in a theory in which physical and subjective probabilities are mixed together. I shall not digress to discuss the philosophy of kinds of probability. (See, for example, Kemble, 1941; Good, 1959; 1965, Chapter 2.) It won't affect what I say whether you believe in the real existence of physical (material) probability or whether you regard it as defined in terms of de Finetti's theorem concerning permutable (exchangeable) events.

I shall first explain the three headings leaving most of the elaborations and historical comments until later.

(i) *Hierarchies of physical probabilities.* The meaning of the first heading is made clear merely by mentioning populations, superpopulations, and super-duper-populations, etc. Reichenbach (1934/1949, Chapter 8) introduced hierarchies of physical probabilities in terms of random sequences, random sequences of random sequences, etc.

(ii) *Hierarchies arising in a subjective theory.* Most of the justifications of the axioms of subjective probability assume sharp probabilities or clear-cut decisions, but there is always some vagueness and one way of trying to cope with it is to allow for the confidence that you feel in your judgements and to represent this confidence by probabilities of a higher type.

(iii) *Mixed hierarchies.* The simplest example of a mixed hierarchy is one of two levels wherein a subjective or perhaps logical distribution is assumed for a physical probability. But when there are only two levels it is somewhat misleading to refer to a "hierarchy".

In case (i), Bayes's theorem is acceptable even to most frequentists; see, for example, von Mises (1942). He made the point, which now seems obvious, that if, in virtue of previous experience, something is "known" about the distribution of a parameter θ, then Bayes's theorem gives information about the final probability of a random variable x whose distribution depends on θ. Presumably by "known" he meant "judged uncontroversially". In short he emphasized that a "non-Bayesian" can use Bayes's theorem in some circumstances, a point that was also implicit in Reichenbach's Chapter 8. The point was worth making in 1942 because statisticians had mostly acquired the habit of using Fisherian techniques which nearly always ignore the possibility that there might sometimes be uncontroversial approximate prior distributions for parameters. F.N. David (1949, pp. 71 & 72) even said that Bayes's *theorem* "is wholy fallacious except under very restrictive conditions" and "... at the present time there are few adherents of Bayes' theorem" von Mises (1942, p.157) blew it by saying that the notion that prior probabilities are non-empirical "cannot be strongly enough refuted". He certainly failed to refute them strongly enough to stem the expansion of modern forms of subjectivistic Bayesianism.

Some people regard the *uncontroversial* uses of Bayes's theorem, that is, those uses acceptable to von Mises, as a case of the empirical Bayes method. Others, such as R.G. Krutchkoff, use the expression "empirical Bayes" only for the more subtle cases where the prior is assumed to exist but drops out of the formula for the posterior expectation of θ. It was in this sense that A.M. Turing used the empirical Bayes method for a classified application in 1941. I applied his method with many elaborations in a paper published much later (Good, 1953) which dealt with the population frequencies of species of animals or plants or words. If, in a sample of N animals, there are n_r species each represented r times, we may call n_r the frequency of the frequency r. Of course $\Sigma r n_r = N$. Let q_r be the population probability of such a species. Turing argued that

$$\mathrm{E}(q_r) = \frac{(r+1)n_{r+1}}{\mathrm{N}n_r} \qquad (1)$$

and I modified this formula to $(r+1)n'_{r+1}/(Nn'_r)$ where $n'_1, n'_2, \ldots,$ is a smoothing of $n_1, n_2, \ldots$, and I generalized the argument to give formulae for the moments of the posterior distribution of q_r. It follows that, in another sample of size N, the total expected frequency of the set of species that were each represented r times (in the first sample) is about $(r+1)n'_{r+1}$, not rn_r as would be suggested by a naive application of the method of maximum likelihood. In particular the probability that the next animal or word that you meet will be one that you have not met before is approximately n_1/N and not the maximum likelihood estimate which is zero. The formula (1) was later obtained by Robbins (1956, p. 159) in relation to the almost identical problem of sampling a large collection of Poisson distributions. In fairness to Robbins it should be noted that he had some of the philosophical ideas of the empirical Bayes method in Robbins (1951) though he did not name the method at that time.

Perhaps a statistical argument is not fully Bayesian unless it is subjective enough to be controversial, even if the controversy is between Bayesians themselves. Any subjective idea is bound to be controversial in spite of the expression "de gustibus non disputandum est" (concerning taste there is no dispute). Perhaps most disputes *are* about taste. We can agree to differ about subjective probabilities but controversies arise when communal decisions have to be made. The controversy cannot be avoided, though it may be decreased, by using priors that are intended to represent ignorance, as in the theories of Jeffreys and of Carnap. (Of course "ignorance" does not here mean ignorance about the prior.) All statistical inference is controversial in any of its applications, though the controversy can be negligible when samples are large enough. Some anti-Bayesians often do not recognize this fact of life. The controversy causes difficulties when a statistican is used as a consultant in a legal battle, for few jurymen or magistrates understand the foundations of statistics, and perhaps only a small fraction even of statisticians do. I think the fraction will be large by 2047 A.D.

Now consider heading (ii), in which at least two of the levels are logical or subjective. This situation arises naturally out of a theory of partially ordered subjective probabilities. In such a theory it is not assumed, given two probabilities p_1 and p_2, that either $p_1 \geq p_2$ or $p_2 \geq p_1$. Of course partial ordering requires that probabilities are not necessarily numerical, but numerical probabilities can be introduced by means of random numbers, shuffled cards etc., and then the theory comes to the same thing as saying that there are upper and lower probabilities, that is, that a probability lies in some interval of values. Keynes (1921) emphasized such a theory except that he dealt with logical rather than subjective probabilities. Koopman (1940a, b) developed axioms for such a theory by making assumptions that seemed complex but become rather convincing when you think about them. I think the simplest possible acceptable theory along these lines was given by Good (1950), and was pretty well justified by C.A.B. Smith (1961). (See also Good, 1962.) Recently the theory of partially-ordered probability has often been called the theory of qualitative probability, though I think the earlier name "partially ordered" is clearer. When we use sharp probabilities it is for the sake of simplicity and provides an example of "rationality of type 2" (Good, 1971c).

If you can say confidently that a logical probability lies in an interval (a,b) it is natural to think it is more likely to be near to the middle of this interval than to the end of it; or perhaps one should convert to log-odds to express a clear preference for the middle. (Taking the middle of the log-odds interval is an invariant rule under addition of weight of evidence.) At any rate this drives one to contemplate the notion of a higher type of probability for describing the first type, even though the first type is not necessarily physical. This is why I discuss hierarchies of probabilities in my paper on rational decisions, Good (1952). Savage (1954, p.58) briefly discusses the notion of hierarchies of subjective probabilities, but he denigrates and dismisses them. He raises two apparent objections. The first, which he heard from Max Woodbury, is that if a primary probability has a distribution expressed in terms of secondary probabilities, then one can perform an integration or summation so as to evaluate a composite primary probability. Thus you would finish up with a sharp value for the primary probability after all. (I don't regard this as an objection.) The second objection that he raises is that there is no reason to stop at secondary probabilities, and you could in principle be led to an infinite hierarchy that would do you no good.

In Good, (1950, p. 41) I had said that higher types of probability might lead to logical difficulties but in Good (1952) I took the point of view that it is mentally healthy to think of your subjective probabilities as estimates of credibilities, that is, of logical probabilities (just as it is healthy for some people to believe in the existence of God). Then the primary probabilities might be logical but the secondary ones might be subjective, and the composite probability obtained by summation would be subjective also. Or the secondary ones might also be logical but the tertiary ones would be subjective. This approach does not deny Max Woodbury's point; in fact it might anticipate it. I regard the use of hierarchical chains as a technique helping you to sharpen your subjective probabilities. Of course if the subjective probabilities at the top of the hierarchy are only partially ordered (as they normally would be if your judgements were made fully explicit), the same will be true of the composite primary or type I probabilities after the summations or integrations are performed. Another development of the hierarchical approach in my 1952 paper is in relation to minimax decision functions. Just as these were introduced to try to meet the difficulty of using ordinary Bayesian decisions, one can define a minimax decision function of type II, to avoid using Bayesian decision functions of type II. (The proposal was slightly modified in Good, 1955.) Leonid Hurwicz (1951) made an identical proposal simultaneously and independently. I still stand by the following two comments in my paper: "... the higher the type the woollier the probabilities ... the higher the type the less the wooliness matters provided [that] the calculations do not become too complicated". (The hierarchical method must often be robust, otherwise, owing to the wooliness of the higher levels, scientists would not agree with one another as often as they do. This is why I claimed that the higher wooliness does not matter much.) Isaac Levi (1973, p. 23), says "Good is prepared to define second order probability distributions..., and third order probability distributions over these, etc., until he gets tired". This was funny, but it would be more accurate to say that I stop when the guessed expected utility of going further becomes negative if the cost is taken into account.

Perhaps the commonest hierarchy that deserves the name comes under heading (iii). The primary probabilities, or probabilities of type I, are physical, the secondary ones are more or less logical, and the tertiary ones are subjective. Or the sequence might be: physical, logical, logical, subjective. In the remainder of my paper I shall discuss hierarchies of these kinds.

2. SMALL PROBABILITIES IN LARGE CONTINGENCY TABLES

I used a hierarchical Bayesian argument in Good (1956) (original version rejected in 1953 I am proud to say) for the estimation of small frequencies in a large pure contingency table with entries (n_{ij}). By a *pure* table I mean one for which there is no clear natural ordering for the rows or for the columns. Let the physical probabilities corresponding to the cells of the table be denoted by p_{ij}, and the marginals by $p_{i.}$ and $p_{.j}$. Then the amount of information concerning a row provided by the observation of a column can be defined as $\log\,[p_{ij}/(p_{i.}p_{.j})]$ and it seemed worth trying the assumption that this has approximately a normal distribution over the table as a whole. This turned out to be a readily acceptable hypothesis for two numerical examples that were examined. In other words it turned out that one could accept the loglinear model

$$\log p_{ij} = \log p_{i.} + \log p_{.j} + \epsilon$$

where ϵ has a normal distribution whose parameters can be estimated from the data. (This was an early example of a loglinear model. Note that if ϵ is replaced by ϵ_{ij} and its distribution is not specified, then the equation does not define a model at all.) If then a frequency n_{ij} is observed it can be regarded as evidence concerning the value of p_{ij}, where p_{ij} has a lognormal distribution. Then an application of Bayes's theorem gives a posterior distribution for p_{ij}, even when $n_{ij} = 0$. This seemed to me an interesting example of estimating the probability of an event that had never occurred, but the referee discouraged me from saying this, possibly because it sounded philosophical. As Jimmie Savage once remarked "'Philosophy' is a dirty ten-lettered word". The lognormal distribution was used as a prior for the parameter p_{ij} and the parameters in this distribution would now-a-days often be called hyperparameters. Perhaps this whole technique could be regarded as a non-controversial use of Bayes's theorem. Incidentally, if it is assumed that $p_{ij}/(p_{i.}p_{.j})$ has a Pearson Type III distribution, the estimates turn out to be not greatly affected, so the method appears to be robust. (The calculation had to be iterative and was an early example of the EM method as pointed out by Dempster *et al*, 1977, p. 19.)

3. MAXIMUM LIKELIHOOD/ENTROPY FOR ESTIMATION IN CONTINGENCY TABLES

For ordinary and multidimensional *population* contingency tables, with some marginal probabilities known, the method of maximum entropy for estimating the probabilities in the individual cells leads to interesting results (Good, 1963). [The principle of maximum entropy was interpreted by Jaynes (1957) as a method for selecting prior distributions. Good (1963) interprets it as a method for formulating hypotheses; in the application it led to hypotheses of vanishing interactions of various orders. Barnard mentions that an early proposer of a principle of maximum entropy was Jean Ville in the Paris conference on the history and philosophy of science in 1949 but I have not yet been able to obtain this reference.] When there is a sample it is suggested in Good (1963, p. 931) that one might find the estimates by maximizing a linear combination of the log-likelihood and the entropy, that is, in the two-dimensional case, by maximizing an expression of the form $\Sigma(n_{ij} - \lambda p_{ij}) \log p_{ij}$, subject to constraints if the marginal probabilities are assumed. [Here (n_{ij}) is the sample and (p_{ij}) the population contingency table.] This technique could be adopted by a non-Bayesian who would think of λ as a "procedure parameter". A Bayesian might call it a hyperparameter because the ML/E method, as we may call it, is equivalent to the maximization of the posterior density when the prior density is proportional to $\Pi p_{ij}^{-\lambda p_{ij}}$. This method has been investigated by my ex-student Pelz (1977). I believe that the best way to estimate the hyperparameter λ is by means of the method of cross-validation or predictive sample reuse, a method that could also be used for comparing the ML/E method with other methods (Good, 1979c). We intend to try this approach.

4. MULTINOMIAL DISTRIBUTIONS

Some hierarchical models that have interested me over a long period are concerned with multinomials and contingency tables, and these models received a lot of attention in my monograph on the estimation of probabilities from a Bayesian point of view (Good, 1965). (See also Good, 1964.) To avoid controversy about purely mathematical methods I there used the terminology of distributions of types I, II and III, without committing myself about whether the probabilities were physical, logical, or subjective. But, in a

Bayesian context, it might be easiest to think of these three kinds of probability as being respectively of the types I, II and III. My next few hundred words are based on Good (1965) where more details can be found although the present discussion also contains some new points.

The estimation of a binomial parameter dates back to Bayes and Laplace, Laplace's estimate being known as "Laplace's law of succession" This is the estimate $(r + 1)/(N + 2)$, where r is the number of successes and N the sample size. This was the first example of a shrinkage estimate. It was based on the uniform prior for the binomial parameter p. The more general conjugate prior of beta form was proposed by the actuary G.F. Hardy (1889). De Morgan (1847) (cited by Lidstone, 1920) generalized Laplace's law of succession to the multinomial case where the frequencies are (n_i) $(i = 1, 2,\ldots,t)$. (I have previously attributed this to Lidstone.) De Morgan's estimate of the i^{th} probability p_i was $(n_i + 1)/(N + t)$ which he obtained from a uniform distribution of $(p_1, p_2,\ldots,p_t)$ in the simplex $\Sigma\, p_i = 1$ by using Dirichlet's multiple integral. The estimate is the logical or subjective expectation of p_i and is also the probability that the next object sampled will belong to the i^{th} category. The general Dirichlet prior, proportional to $\Pi p_i^{k_i-1}$, leads to the estimate $(n_i + k_i)/(N + \Sigma k_j)$ for p_i. But if the information concerning the t categories is symmetrical it is adequate, at the first Bayesian level, to use the prior proportional to Πp_i^{k-1} which leads to the estimates $(n_i + k)/(N + tk)$. In fact we can formulate the Duns-Ockham hyper-razor as "What can be done with fewer (hyper)parameters is done in vain with more". ("Ockham's razor" had been emphasized about twenty years before Ockham by the famous medieval philosopher John Duns Scotus.) We can regard k both as a flattening constant or as the hyperparameter in the symmetric Dirichlet prior. The proposal of using a continuous linear combination of Dirichlet priors, symmetric or otherwise, occurs in Good (1965, p.25). Various authors had previously proposed explicitly or implicitly that a single value of k should be used but I am convinced that we need to go up one level. (Barnard tells me he used a combination of two beta priors in an unpublished paper presented at a conference in Bristol in about 1953 because he wanted a bimodal prior.)

The philosopher W.E. Johnson (1932) considered the problem of what he called "multiple sampling", that is, sampling from a t-letter alphabet. He assumed permutability of the N letters of the sample (later called "exchangeability" though "permutability" is a slightly better term). Thus he was really considering *multinomial sampling. He further assumed* what I call his "sufficientness postulate", namely that the credibility (logical probability) that the next letter sampled will be of category i depends only on n_i, t, and N, and does not depend on the ratios of the other $t - 1$ frequencies. Under these assumptions he proved that the probability that the next letter sampled will be of category i is $(n_i + k)/(N + tk)$, but he gave no rules for determining k. His proof was correct when $t \geq 3$. He was presumably unaware of the relationship of this estimate to the symmetric Dirichlet prior. The estimate does not merely follow from the symmetric Dirichlet prior; it also implies it, in virtue of a generalization of de Finetti's theorem. (This particular generalization follows neatly from a purely mathematical theorem due to Hildebrandt and Schoenberg; see Good, 1965, p. 22.) De Morgan's estimate is the case $k = 1$. Maximum Likelihood estimation is equivalent to taking $k = 0$. The estimates arising out of the invariant priors of Jeffreys (1946) and Perks (1947) correspond to the flattening constants $k = ½$ and $k = 1/t$.

Johnson's sufficientness assumption is unconvincing because if the *frequencies* $n_2, n_3,\ldots,n_t$ *are far from equal it would be natural to believe that* p_1 is more likely to be far from $1/t$ than if $n_2, n_3,\ldots,n_t$ are nearly equal. Hence it seemed to me that the "roughness" of the frequency count (n_i) should be taken into account. Since roughness can be measured by a scalar I felt that k could be estimated from the sample (and approximately from its roughness), or alternatively that a hyperprior could be assumed for k, say with a density function $\phi(k)$. This would be equivalent to assuming a prior for the p_i's, with density

$$\int_0^\infty \frac{\Gamma(tk)\Pi p_i^{k-1}\phi(k)dk}{[\Gamma(k)]^t}$$

Those who do not want to assume a hyperprior could instead estimate k say by Type II Maximum Likelihood or by other methods in which the estimate of k is related to $X^2 = \frac{t}{N}\Sigma\, (n_i\text{-}N/t)^2$. These methods were also developed by Good (1965, 1966, 1967). Good (1967) was mainly concerned with the Bayes factor, provided by a sample (n_i), against the null hypothesis $p_i = 1/t$ $(i = 1, 2,\ldots,t)$. The estimation of the cell probabilities p_i was also covered. (It seems to me to be usually wrong in principle to assume distinct priors, given the non-null hypothesis, according as you are doing estimation or significance testing, except that I believe that more accurate priors are required for the latter purpose.) The null hypothesis corresponds to the complete flattening $k = \infty$ and we may denote it by H_∞. Let H_k denote the non-null hypothesis that the prior is the symmetric Dirichlet with hyperparameter k. Let $F\,(k)$ denote the Bayes factor in favour of H_k as against H_∞, provided by a sample (n_i). (See Good, 1957, p. 862; or 1967, p. 406.) If k has a hyperprior density $\phi(k)$, then the Bayes factor F against H_∞ is

$$F = \int_0^\infty F\,(k)\phi(k)dk\ ,$$

$\phi(k)$ must be a proper density, otherwise F would reduce to 1, in other words the evidence would be killed. This is an interesting example where impropriety is a felony. One might try to be noncommittal about the value of k and the usual way of being noncommittal about a positive parameter k is to use the Jeffreys-Haldane density $1/k$ which is improper. This can be approximated by the log-Cauchy density which has the further advantage that its quantiles are related in a simple manner to its hyperhyperparameters (Good, 1969, pp. 45-46). One can determine the hyperhyperparameters by guessing the upper and lower quartiles of the repeat rate Σp_i^2, given the non-null hypothesis, and thereby avoid even a misdemeanour. The Bayes factor F is insensitive to moderate changes in the quartiles of the log-Cauchy hyperprior, and the estimates of the p_i's are even more robust. If you prefer not to assume a hyperprior then a type II or second order or second level Maximum Likelihood method is available because $F(k)$ has a unique maximum F_{max} if $X^2 > t - 1$. This was conjectured by Good (1965, p. 37) largely proved by Good (1975) and completely proved by Levin and Reeds (1977). Other methods of estimating k are proposed by Good (1965, pp. 27, 33, 34) and by Bishop, Fienberg and Holland (1975, Chapter 12). When a hyperparameter is estimated the latter authors call the method "pseudo-Bayesian". It is an example of a Bayes/non-Bayes compromise.

F_{max} is an example of a Type II (or second order or second level) Likelihood Ratio defined in terms of the hyperparametric space which is one-dimensional. Hence the asymptotic distribution of F_{max} is proportional to a chi-squared with one degree of freedom. In 1967 the accuracy of this approximation was not known but it was found to be fairly accurate in numerous examples in Good and Crook (1974), even down to tail-area probabilities as small as 10^{-16}. We do not know why it should be an adequate approximation in such extreme tails.

5. INDEPENDENCE IN CONTINGENCY TABLES

Good (1965) began the extension of the multinomial methods to the problem of testing independence of the rows and columns of contingency tables, and this work was continued in Good (1976a) where extensions to three and more dimensions were also considered. But I shall here consider only ordinary (two-dimensional) tables with r rows and s columns. The case $r = s = 2$ is of special interest because 2×2 tables occur so frequently in practice.

As is well known outside Bayesian statistics, there are three ways of sampling a contingency table, known as sampling Models 1, 2 and 3. In Model 1, sampling is random from the whole population; in Model 2, the row totals (or the comumn totals) are fixed in advance by the statistician; and in Model 3 both the row and column totals are fixed. Model 3 might seem unreasonable at first but it can easily arise. Denote the corresponding Bayes factors against the null hypothesis H of independence by F_1, F_2, and F_3. But in our latest model it turns out that $F_1 = F_2$ because in this model the fixing of the row totals alone provides no evidence for or against H. The model also neglects any evidence that there might be in the order of rows or of columns; in other words we restrict our attention in effect to "pure" contingency tables. This is of course, also done when X^2 or the likelihood-ratio statistic is used.

The basic assumption in the analysis is that, given the non-null hypothesis $\bar{H}$, the prior for the physical probabilities p_{ij} in the table is a mixture of symmetric Dirichlet's. (Previous literature on contingency tables had discussed symmetric Dirichlet distributions but not mixtures.) From this assumption F_1 and F_3 can be calculated. We can deduce FRACT (the factor against H provided by the row and column totals alone, in Model 1) because FRACT $= F_1/F_3$. A large number of numerical calculations have been done and were reported in Crook and Good (1980). We found that FRACT usually lies between ½ and 2½ when neither of the two sets of marginal totals is very rough and the two sets are not both very flat, and we gave intuitive reasons for these exceptions. We did not report the results for 2×2 tables in that paper but we have done the calculations for this case with the sample size $N = 20$. We find, for example, with our assumptions, that FRACT $= 1.48$ for the table with margins [5,15;7,13]; FRACT $= 2.53$ for [10,10;10,10]; FRACT $=$ 2.56 for [1,19;2,18]; and FRACT $= 8.65$ for the extreme case [1,19;1,19].

If the mixture of Dirichlet's is replaced by a single symmetrical Dirichlet with hyperparameter k, then F_3 is replaced by $F_3(k)$, and $\max_k F_3(k)$ is a Type II Likelihood Ratio. Its asymptotic distribution again turns out to be fairly accurate in the extreme tail of the distribution, even down to tail-area probabilities such as 10^{-40}. The unimodality of $F_3(k)$ when $X^2 > (r - 1)(s - 1)$ has yet to be proved, but is well supported by our numerical results.

I noticed only as recently as May 1978 that the consideration of contingency tables sheds light on the hyperprior ϕ for multinomials. This was first reported in Good (1979b). We write $\phi(t,k)$ instead of $\phi(k)$ to indicate that it might depend on t as well as k. The prior for a t-category multinomial is then $D^*(t)$ where

$$D^*(t) = \int_0^\infty D(t,k)\phi(t,k)dk$$

and where $D(t,k)$ denotes the symmetric Dirichlet density. Our assumption of $D^*(rs)$, given $\bar{H}$ and Model 1, implies the prior $\int_0^\infty D(r,sk)\phi(rs,k)dk$ for the row totals. But, if the row totals alone contain no evidence concerning H, this must be mathematically independent of s and it can be deduced that $\phi(t,k)$ must be of the form $\psi(tk)/k$. Strictly therefore some of the calculations in

Good and Crook (1974) should be repeated, but of course the distribution of the Type II Likelihood Ratio is unaffected, and we have reason to believe the remaining results are robust. This example shows how logical arguments can help to make subjective probabilities more logical. Logical probabilities are an ideal towards which we strive but seldom attain.

A spin-off from the work on contingency tables has been the light it sheds on the classical purely combinatorial problem of the enumeration of rectangular arrays of integers (Good and Crook, 1977; Good, 1979a). This problem had not previously been treated by statistical methods as far as I know.

T. Leonard has used hierarchical Bayesian methods for analyzing contingency tables and multinomial distributions, but since he has attended this conference I shall leave it to him to reference his work in the discussion of the present paper.

6. PROBABILITY DENSITY ESTIMATION AND BUMP HUNTING

Probability density estimation has been a popular activity since at least the nineteenth century, but bump-hunting, which is closely related to it, is I think comparatively recent. There is a short discussion of the matter in Good (1950, pp. 86-87) where the "bumpiness" of a curve is defined as the number of points of inflexion, though half this number is a slightly better definition. The number of bumps was proposed as one measure of complexity, and the greater the number the smaller the initial probability of the density curve *ceteris paribus*.

In the 1970 Waterloo conference, Orear and Cassel (1971) said that bump-hunting is "one of the major current activities of experimental physicists". In the discussion Good (1971a) suggested the idea of choosing a density function f by maximizing $\Sigma \log f(x_i) - \beta R$, that is, log-likelihood minus a roughness penalty proportional to a measure R of roughness of the density curve. (Without the penalty term one gets $1/N$ of a Dirac function at each observation.) It was pointed out that the problem combines density estimation with significance testing. In Good (1971b) the argument is taken further and it is mentioned that $\exp(-\beta R)$ can be regarded as the prior density of f in function space. In this Bayesian interpretation β is a hyperparameter. (There were 21 misprints in this short article, owing to a British dock strike.) The method was developed in considerable detail by Good and Gaskins (1971, 1972) and applied to two real examples, one relating to high-energy physics and the other to the analysis of chondrites (a common kind of meteorite containing round pellets) by Good and Gaskins (1979). In the latter work, the estimation of the hyperparameter was made by means of non-Bayesian tests of goodness of fit so as to avoid controversies arising out of the use of hyperpriors.

Leonard (1978, p. 129) mentions that he hopes to report a hierarchical form of his approach to density estimation. Also he applies his method to the chondrite data, and he brought this data to my attention so that our methods could be compared.

7. INFERENCE ABOUT NORMAL DISTRIBUTIONS AND LINEAR MODELS

In 1969 I suggested to my student John M. Rogers that he might consider analogies of the multinomial work for the estimation of the parameters of multivariate normal distributions. It turned out that even the univariate problems were complicated and he completed his thesis without considering the multivariate problems. He considered the estimation of *a* (univariate) normal mean when the prior contains hyperparameters. The priors were of both normal and Cauchy form (Rogers, 1974) and the hyperparameters were estimated by type II maximum likelihood.

Meanwhile hierarchical Bayesian models with three or four levels or stages had been introduced for inferences about normal distributions and linear models by Lindley (1971) and by Lindley and Smith (1972.) A survey of these matters could be better prepared by Lindley so I shall say no more about them.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health, Grant Number R01 GM18770.

REFERENCES

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis* Harvard, Mass: M.I.T. Press.

CROOK, J.F. and GOOD, I.J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, II. *Ann. Statist.* (to be published).

DAVID, F.N. (1949). *Probability Theory for Statistical Methods*. Cambridge: University Press.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39**, 1-38 (with discussion).

GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

— (1952). Rational decisions. *J.Roy Statist. Soc. B* **14**, 107-114

— (1953). On the population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237-264.

— (1955). Contribution to the discussion on the Symposium on Linear Programming. *J. Roy. Statist. Soc. B.* **17**, 194-196.

— (1956). On the estimation of small frequencies in contingency tables. *J. Roy. Statist. Soc. B,* **18**, 113-124.

— (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28**, 861-881.

— (1959). Kinds of probability. *Science* **127**, 443-447.

— (1962). Subjective probability as the measure of a non-measurable set. In *Logic, Methodology and Philosophy of Science* (Nagel, E., Suppes, P., and Tarski, A. eds), 319-329.

— (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist,* **34**, 911-934.

— (1964). Contribution to the discussion of A.R. Thatcher Relationships between Bayesian and confidence limits for predictions. *J. Roy. Statist. Soc. B,* **26**, 204-205.

— (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Harvard, Mass: M.I.T. Press.

— (1966). How to estimate probabilities. *J. Inst. Math. Applics.* **2**, 364-383.

— (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. B* **29**, 399-431.

— (1969). A subjective analysis of Bode's law and an 'objective' test for approximate numerical rationality. *J. Amer. Statist. Assoc.* **64**, 23-66 (with discussion).

— (1971a). Contribution to the discussion of Orear and Cassel (1971), 284-286.

— (1971b). Nonparametric roughness penalty for probability densities. *Nature Physical Science* **229**, 29-30.

— (1971c). Twenty-seven principles of rationality. In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott. ed.) 123-127, Toronto: Holt, Rinehart and Winston.

— (1975). The Bayes factor against equiprobability of a multinomial population assuming a symmetric Dirichlet prior. *Ann. Statist.* **3**, 246-250.

— (1976a). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**, 1159-1189.

— (1976b). The Bayesian influence or how to sweep subjectivism under the carpet. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* **2** (C.A. Hooker and W. Harper, eds.) 125-174, Dordrecht, Holland: D. Reidel.

— (1979a). A comparison of some statistical estimates for the numbers of contingency tables, item C26 in "Comments, Conjectures, and Conclusions". In *J. Statist. Comput. Simul.* **8**, 312-314.

— (1979b). The contributions of Jeffreys to Bayesian statistics. In *Studies in Bayesian Econometrics and Statistics in Honor of Harold Jeffreys.* (A. Zellner, ed.), 21-34. Amsterdam: North Holland.

— (1979c). Predictive sample reuse and the estimation of probabilities. *J. Statist. Comput. Simul.* **9**, 238-239.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

— (1977). The enumeration of arrays and a generalization related to contingency tables. *Discrete Mathematics* **19**, 23-45.

GOOD, I.J. and GASKINS, R.A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255-277.

— (1972). Global nonparametric estimation of probability densities. *Virginia J. of Science* **23**, 171-193

— (1979). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 42-73 (with discussion).

HARDY, G.F. (1889). In correspondence in Insurance Record. Reprinted in *Trans. Fac. Actuaries* **8** (1920), 174-182.

HURWICZ, L. (1951). Some specification problems and applications to econometric models, *Econometrics* **19**, 343-344 (abstract).

JAYNES, E.T. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106**, 620-630.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. (London), A.* **186**, 453-461.

JOHNSON, W.E. (1932). Appendix (ed. R.B. Braithwaite) to Probability: deductive and inductive problems. *Mind* **41**, 421-423.

KEMBLE, E.C. (1941). The probability concept. *Philosophy of Science* **8**, 204-232.

KEYNES, J.M. (1921). *A Treatise on Probability.* London: Macmillan.

KOOPMANM, B.O. (1940a). The basis of probability. *Bull. Amer. Math. Soc.* **46**, 763-764.

— (1940b). The axioms and algebra of intuitive probability. *Ann. Math.* **41**, 269-292.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc, B,* **40**, 113-146 (with discussion).

LEVI, I. (1973). Inductive logic and the improvement of knowledge. *Tech. Rep,* Columbia University.

LEVIN, B. and REEDS, J. (1977). Compound multinomial likelihood functions: proof of a conjecture of I.J. Good., *Ann. Statist.* **5**, 79-87.

LIDSTONE, G.J. (1920). Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Trans. Fac. Actuar.* **8**, 182-192.

LINDLEY, D.V. (1971). The estimation of many parameters. In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott; eds.) 435-455, (with discussion). Toronto: Holt, Rinehart and Winston.

LINDLEY, D.V. and SMITH, A.F.M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. B.* **34**, 1-41 (with discussion).

DE MORGAN, A. (1847). Theory of probabilities. *Encyclopaedia Metropolitana*, **2**, 393-490.

OREAR, J. and CASSEL, D. (1971). Applications of statistical inference to physics. In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott. eds.) 280-288 (with discussion). Toronto: Holt, Rinehart and Winston.

PELZ, W. (1977). *Topics on the estimation of small probabilities.* Ph D thesis, Virginia Polytechnic Institute and State University.

PERKS, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuar* **73**, 285-312.

REICHENBACH, H. (1949). *The Theory of Probability.* Berkeley: University of California Press.

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. 2nd Berkeley Symp.* 131-148. Berkeley: University of California Press.

— (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.* **1**, 157-163. Berkeley: University of California Press..

ROGERS, J.M. (1974). *Some examples of compromises between Bayesian and non-Bayesian statistical methods.* Ph. D. Thesis, Virginia Polytechnic Institute and State University.

SAVAGE, L.J. (1954). *The Foundations of Statistics.* New York: Wiley.

SMITH, C.A.B. (1961). Consistency in statistical inference and decision. *J. Roy. Statist. Soc. B.* **23**, 1-37 (with discussion).

VON MISES, R. (1942). On the correct use of Bayes's formula. *Ann. Math. Statist.* **13**, 156-165.

DISCUSSION

M.H. DE GROOT (*Carnegie-Mellon University*):

I am very glad that top-notch statisticians like Professor Dickey are devoting serious effort to the important problem of assessing subjective probabilities. As Bayesian theory becomes the prevalent methodology for public decision making, it is essential that we learn how to elicit an expert's information and opinions in terms of his probabilities.

However, I do have some fundamental questions about the philosophy and foundations that underly this paper, and some of the other papers that Professor Dickey cites. It seems to me that there is a fundamental contradiction in the idea of specifying a probability model for the inconsistencies that are present in the aspects that the assessor specifies. If assessors are not consistent in assessing their subjective probabilities or aspects, then the axioms of subjective probability are not satisfied and probability, as I understand it, does not exist. How then can we speak of a probability model for the assessor's specifications? Must we think of the world as being divided into consistent fitters (or statisticians) and inconsistent assessors?

What are the underlying "true" subjective probabilities that the statistician is trying to fit? Do we really have to assume that there is a coherent little man or woman inside each of us, crying out weakly to be heard? Or that we can split our personality and have our coherent self monitor our incoherent self?

The trend in statistics is distinctly toward the prediction of observables and away from the estimation of unknown parameters. Are we not moving backwards when we try to estimate unknowable and perhaps non-existent true probabilities?

Professor Dickey states clearly that he is treating only one aspect of a complicated, recursive assessment process. In its full development this process must necessarily include continuous internal checking and looping backwards to reconcile the specified aspects, not only among themselves, but with a myriad of other related assessments as well.

How do we know that the values we reach by following such a process have any meaning at all? How do we know that if another statistician had done it all over again with the same assessor in the morning instead of the afternoon, he would have arrived at the same values? What mechanism is there to recognize and handle non-convergence? A related, more technical, question is whether *any* member of the particular parametric family that Professor Dickey uses in a particular problem fits an assessor's specifications satisfactorily. The answer would seem to depend on the uses to which the assessments are to be put.

Professor Dickey is probably his own best discussant since he is clearly aware of all these difficulties. I mention them here not to minimize the substantial accomplishments of this complex paper, but to emphasize the long way we still have to go to make probability assessments a meaningful reality.

It is a pleasure for me to take this opportunity to acknowledge and thank Professor Good for his creative and substantial contributions, not only to the theory of hierarchical models, but to so much of Bayesian statistics. I strongly agree with Professor Good on the importance and usefulness of hierarchical models. However, I do not see any need to consider physical, logical, and subjective probabilities as being different types of probabilities. Subjective probabilities would seem to be enough; they unify the theory and are usually convenient to use.

Even when dealing with "physical" probabilities, it is not necessary to consider populations, superpopulations, and (super)^n populations. The prototype of a hierarchical model is a discrete time Markov process. The probability distribution of the state X_t of the process at a given time (the first-level distribution) depends on the state X_{t-1} at the previous time, and the distribution of X_{t-1} (the second-level distribution) depends in turn on the state X_{t-2}, etc.

In this paper, Professor Good emphasizes the problem of making inferences about the first-level probabilities, but one reason for using a hierarchical model is because we are interested in learning about the higher-level probabilities as well [see, e.g., Lindqvist (1977, 1978)]. Thus, on the basis of our observation of the current state X_t of a Markov process, we might very well be interested in making inferences about the states of the process $X_{t-1}, \ldots, X_{t-k}$ at the k previous stages. For example, having observed my own height, I might be interested in making inferences about the heights of my father, his father, and his grandfather.

Professor Good states that the higher the level in the hierarchy "the woollier the probabilities" and "the less the woolliness matters". I think that I agree with these statements, but I confess that I am not sure what woolly probabilities are. (I suppose that Poisson probabilities are woolly because their mean is lambda.) A somewhat related question has been considered by Goel and me (1979). In that paper, we are interested in learning about the parameters that enter at various levels of a hierarchical model. We show that in terms of several different standard measures of uncertainty and information, the expected reduction in uncertainty and expected gain in information decreases as we move through higher levels of the hierarchy away from the observations themselves.

I have somewhat mixed feeling about the Bayes/non-Bayes compromise that Professor Good proposes. I do not believe that it provides a sound basis for a philosophy of statistics. It does not yield a unified theory and cannot generate fundamental general principles. I do not believe that higher-level prior distributions are more controversial than first-level priors, and I do not believe that in principle there is any reason to estimate hyperparameters by non-Bayesian methods. On the other hand, a Bayes/non-Bayes compromise sounds like a reasonable way to proceed in any particular problem. The difficulty lies in trying to reach a reasonable and widely acceptable compromise in each special case. Thus, if a Bayes/non-Bayes compromise is our destiny, as it very well might be, then I believe that we will unfortunately continue to see a wide gap between statistical theory and statistical practice.

M.R. NOVICK (*The University of Iowa*):

The papers of I.J. Good and J.M. Dickey are both helpful contributions to Bayesian theory and methodology. They differ, however, in several respects. The Good paper is offered as a history of the personal contributions of its author to hierarchical Bayesian methodology. In fact, it provides a useful historical overview and integration in this area. The Dickey paper purports to offer a theory for stochastic assessment, but, in fact, the contribution is primarily mathematical.

Professor Good begins by letting us peer into his crystal ball which shows a future in which all of statistical methodology is based on methods that are either Bayesian, empirical Bayesian, or some combination of, or compromise between, the two sets of techniques. My own crystal ball provides a similar prevision.

This conjectured system of the future apparently makes use of the concepts of frequency probabilities (or propensities), logical probabilities, and subjective probabilities. it emphasizes the Bayesian hierarchical model and interestingly suggests that the various kinds of probabilities can find use at a variety of levels in the hierarchy. I must say that I like this prevision and hope that Professor Good and I are correct.

It would have been nice to have had some brief discussion of the relationship between these various kinds of probabilities, but we ought, I think, to be thankful for what Professor Good has given us. To supplement this I think that a reference or two might be useful. For my own part, I have always found Rudolf Carnap's treatment of the two kinds of probabilities to be very useful. His distinction between, and indication of the relationship between frequency probabilities (or propensities) and logical probabilities I find very useful. With respect to the relationship between subjective probabilities and logical probabilities, I have always found Jimmy Savage's treatment to be very helpful. His simple description of the relationship of these two was to assert that logical probability was a limiting case of subjective probability when so many constraints were placed on the system to permit the assessor no judgment in setting the subjective probabilities. I should also like to point to work by D.V. Lindley and myself relating propensities to subjective probabilities and each of these to the concept of exchangeability which is the central concept of the hierarchical models.

With respect to the relationship between the Bayesian hierarchical models and the so-called empirical Bayes models, I'm not sure that Professor Good's characterization and mine will be identical. Let me offer mine with the understanding that Professor Good will have an opportunity to present his own.

Jimmy Savage used to talk about statistical rhetoric and often pointed out the gamesmanship employed by classical statisticians when they named some of their estimates as unbiased. How, he would say, would it be possible for anyone to use an estimate that was called biased? Fortunately, we are no longer bothered by this rhetoric and most of us routinely use "biased" estimates in our work. I must say, however, that I wish our textbook writers would be more forthright and remove this objectionable terminology from statistical theory. I would also wish that pseudo-Bayesians would cease referring to their work as empirical Bayes. There is really nothing more empirical about their work than any other statistical work using hierarchical models. In fact, one could argue that from a Bayesian point of view some empirical Bayesian methods involve a mild and possibly very acceptable form of cheating, which is to say the use of a conditional rather than a marginal distribution, conditioning on a statistic obtained from the sample at hand. Now we all know that some robustness arguments can lead us to accept the use of such a conditional distribution instead of a marginal distribution, but it certainly must be judged a bit bold to label this method empirical.

In saying this I by no means suggest than these pseudo-Bayesian methods will always lead to poor results. Quite the contrary is true. Often they will lead to very good results and sometimes they will lead to better results than are obtained using fully Bayesian methods when, in the latter case, care is not exercised in the choice of the final stage prior distribution.

One might think that there is a contradiction here in that using pseudo-Bayesian methods leads to good results and using proper methods sometimes leads to poor results. The reason for this is that strictly Bayesian methods are very difficult to employ and sometimes very sensitive because they require the assessment of certain probabilities that are, indeed, very difficult to assess. As a result of this the assessments may not be made and so-called flat or indifference type prior distributions may be used. As it turns out this may lead to very poor results. However, when reasonably good methods are available for assessing the parameters at the last stage of the Bayesian model, we can expect very satisfactory results indeed.

Another disadvantage of pseudo-Bayesian methods is that they provide only point estimates of parameters or variables and, therefore, lead to a second imprecision that is typically found in the application of classical methods.

If we look at the typical introductory book on classical statistics and study the section on regression and prediction carefully, we see the process as one of point estimation of the intercept, the regression coefficients and the residual variance. When this estimation is completed the inference by the reader is that these estimated values should be used as the true values of the parameters and that they should, therefore, be inserted in the model and predictions made with the model using the estimated residual variance to give the error of prediction. In fact, this is incorrect and we know that the correct procedure is to compute a confidence interval for an $(n+1)st$ observation conditioned on the previously observed (x,y) pairs. This yields the same point estimate, but provides for a larger standard error than the simpler procedure. The same difficulty arises with empirical Bayes procedures. Whereas in a strictly Bayesian method we automatically derive the conditional distribution of the dependent variable given the independent variables and marginalized with respect to the posterior distribution on the parameters. If you do Bayesian statistics carefully you get the right answer, but you must be more careful than with pseudo-Bayesian methods.

My crystal ball shows that we shall be looking for and finding better ways to fit prior distributions for the final level in the hierarchical model and when this proves impossible we shall condition where necessary and, as a last resort, we may condition on observed sample values. We will, however, recognize such conditioning for what it is and treat our results accordingly.

The problem of assessing a reasonable prior distribution for the final stage in a hierarchical model is a difficult one, and one that typicaly cannot be avoided. I cannot offer a method that will work in all situations, but a very simple method is available in some cases.

Consider the true long run proportion of correct responses, π_i, associated with person i for $i = 1,2,\ldots,m$. Let the observations for each person be binomial (n, π_i) and let the persons be exchangeable to You. Then if I assess your prior distribution for π_i, coherence will demand that the mean and variance of your distribution for π_i will reflect your belief about the mean and variance of the *distribution of the* π_i*'s*. While this coherence argument is usually taken in the opposite direction, it is equally valid in the stated direction and this may be helpful in assessing a prior distribution at the second level of the hierarchy.

Professor Dickey's paper troubles me. It is not that I do not value his contribution. On the contrary, I think that some of the work may be very useful. But I am troubled by the title and by the implication that the work in this paper somehow *solves* any problem in the assessment of subjective probabilities. It does not. What it does is to provide us with some mathematical models and some mathematical results which should prove very useful in attempting a subjective probability assessment. It, in itself, does nothing to provide for that assessment. The same may be said, though with slightly less force about the cited paper by Kadane, Dickey, Winkler, Smith and Peters.

The original version of the Dickey paper contained no integration with or reference to the psychological or psychometric literature. The final version does not correct this error. I had hoped for more. Bayesian statistics, in its complete form, involves the use of prior probabilities and utilities. The assessment of these beliefs and values involves a psychometric process as important to Bayesian statistics as any mathematical work. A rich psychometric literature should not be ignored. A recent technical report by Gokhale and Press (1979) would seem to set a more useful example than either the Dickey or Kadane, et. al. papers however important the mathematical contributions may be in the two cases. Both papers would be improved by more accurate titling.

S. GEISSER (*University of Minnesota*):

I would like to reinforce Professor Good's point on using predictive sample reuse techniques for the estimation of hyperparameters. Examples of its use in this regard appear in Geisser (1975a, 1975b).

It also appears to me that there are many more ways than three to sample a contingency table. Various combinations of cells in a table can just as easily be fixed. For example, instead of having double binomials in a 2 x 2 table, one could just as easily have double negative binomials or combinations there of. This could typically occur in trials which were generated squentially.

T. LEONARD (*University of Warwick*):

My midnight game of chess with Professor Good has, for me, been one of the highlights of this conference (ranking with some enjoyable discotheque visits) and has confirmed my impression that Bayesianism is fortunate to possess and adherent with such wide-ranging intellectual abilities. I would however like to express the opinion that the razor dictum discussed by Professor Good may not be completely appropriate, in the situation he discusses. By reducing the number of hyperparameters we might even obtain posterior estimates with properties which are different in spirit to those which we were anticipating. To give three examples:

1.- When estimating multinomial probabilities, Good rightly avoid Johnson's sufficientness postulate in order to allow for possible histogram smoothness. He however restricts himself to a single hyperparameter (referred to by Fienberg and Holland as the flattening constant) and thereby flattens the proportions rather than smoothing them by allowing for the ordering of the cells of the histogram. In a previous paper (Leonard, 1973) I showed that by using two hyperparameters (one for shrinking towards a prior estimated and one for smoothing) we could obtain the sorts of properties which Good was anticipating during his discussion of Johnson's sufficientness postulate. Formulations like this would of course be difficult using the Dirichet distribution, because of its highly specialised covariance structure.

2.- Professor Good has contributed a number of interesting papers to the areas of two-way and multi-way contingency tables. However, he usually contents himself with a single hyperparameter, and assumes a symmetric Dirichlet prior across the whole table. This immediately leads to posterior estimates which flatten the cell probabilities regardless of such aspects as row and column dependence; alternative prior structures are discussed by Leonard (1975).

3.- The paper by Good and Gaskins in 1971 provided one of the pioneering contributions to the field of density estimation. However, he again constrains himself to a single hyperparameter which plays the role of the flattening constant in multinomial problems. Since there is no further hyperparameter the practical viability of the procedure is again restricted. Good's estimates again flatten much more than they smooth.

A great deal of interesting work has been carried out on the complex marginal likelihood of the hyperparameter under multinomial-Dirichlet assumptions. It is nice that the unimodality of this likelihood has been proved and that so many of its topological properties are now known. An alternative may be obtained by referring to an approximation to the marginal distribution of the chi-squared statistic. This yields a simple approximation to the marginal likelihood of the flattening constant whilst preserving its important topological properties. Details are described by Leonard (1977).

In my opinion a formally Bayesian analysis of the multinomial Dirichlet problem may be a bit more sensitive to the choice of hyper-prior than Professor Good might think. The tails of the marginal likelihood of the flattening constant do not approach zero at infinity; they either approach infinity or move downwards towards a positive asymptote. Therefore, given any hyperprior, an arbitrarily small wiggle taken arbitrarily far out in the tails can always be performed in such a way that the corresponding change in the hyperposterior will substantially affect the posterior distribution of the cell probabilities (which involved an integration over the whole range of the hyperposterior). This difficulty does not seem to be apparent under a multivariate logit/multivariate normal prior approach (e.g. Leonard 1973.)

D.V. LINDLEY (*University College London*):

Is life as complicated as Good suggests? An alternative view is that there is only *one* type of probability, expressing Your coherent view of the world; and one type of rationality, coherence. (Of course, there is chance, as well.) What Good calls subjective probability may not be probability at all, if it is what a *real* subject believes: for a real subject is incoherent, that is, does not obey the rules of probability (see DeGroot's paper.) Similarly, the important hierarchical model of Good's is only a way of modelling one's coherent view of the world. As soon as one introduces parameters, one has moved way from observables: and hyperparameters describe parameters just as parameters describe observables. There is a unity here that Good's diversity seems to undermine. Of course, we have to use simple methods — but these should be viewed as approximations to a fully coherent view.

REPLY TO THE DISCUSSION

J.M. DICKEY (*University College of Wales, Aberystwyth*):

In presenting this paper, I asked the audience to forget a few of their favourite things:

1. Sampling models
2. Likelihoods
3. Bayes' theorem
4. Prior and posterior distributions of parameters.

The idea of statistical data was also discarded initially, and later brought back, in regard to elicited information concerning expert opinions, that is, data about beliefs. All these favourite ideas reappear, of course, in the rather special situation when belief is updated to take *formal* account of *modeled* statistical sample data. But the concern is not merely with helping Bayesian statisticians choose their prior distributions. A whole paradigm has been proposed in which subjective Bayesian inference from modeled statistical data forms just a part.

I have recently used the term "subjective-probability modeling" to emphasize the point that probability models, similar in some ways to the sampling models used by statisticians, can be fitted to describe degrees of belief (and not just belief about parameters). In particular, a *multivariate* opinion can be modeled, together with its functional dependence on concomitant variables. It is this point that I would hope to make above all.

One difficulty with the name "subjective-probability modeling" is that by omitting the hyphen, one could mistakenly read the word "subjective" as an adjective modifying the word "modeling". Whereas, in fact, the modeling process itself is no more nor no less subjective than other mathematical modeling processes. It is just that the object which the model is intended to represent is a person's opinion.

Again, as in any mathematical modeling, never (or hardly ever) is the assumed class of models *exactly true* of the object being modeled. Therefore, Professor DeGroot's question about the satisfactoriness of the particular parametric families put forth in the paper would seem to hinge on the question of their usefulness as experienced in future applications.

I see no "fundamental contradiction" in separating the two concepts of coherence of beliefs and coherence of elicited aspects of beliefs. Coherent beliefs can underlie noncoherent elicitations, just as a physical quantity can underlie the measurements of it. However, conception is one thing, and practice another. In practice, no mathematical model is ever exactly true, and this statement applies as well to the mathematical model of coherence. But in Physics, just because ideal gases do not exist in the world, does not cause us to discard the concept of an ideal gas as useless. (Not only are the *measurements* of volume, pressure, and temperature subject to error, but the underlying *real* volumes, pressures, and temperatures do not follow exactly the laws for an ideal gas). Neither should we abstain in probabilistic opinion-modeling from using the concepts of coherence, and especially since the theory is proposed as a normative theory, rather than a descriptive one.

I concur with Professor DeGroot in his concern over the potential for a destructive dependence of the fit on the fitting strategy. Again, future experience will tell.

I wonder why Prof. Novick has chosen to pretend that I have not proposed assessment algorithms (as in Kadane *et al* 1978) nor criteria for fitting (present paper). It would have been better had he reported, from his extensive experience with interactive computer programs, his views on specific ways in which the proposals are or are not practical. For example, what of my suggestion that errors of elicitation be treated as having variances proportional to the variances in the subjective-probability model? (I note that a recent release of Prof. Novick's computer program system CADA implements the algorithm of Kadane *et al*).

I apologize to Professor Novick for not integrating my paper with the psychological literature. I did not feel that the comments I had in mind on this literature were relevant to the present paper, although the topic of the paper can be considered a psychometric topic. Perhaps, a few words here will go some way toward making amends.

For the most part, the papers I have seen in the psychological literature use the very simplest of subjective-probability models, where the model does not have enough structure to produce interesting derived statements concerning a single experts's opinion on a particular occasion. (See, for example, the second half of Slovic and Lichtenstein 1971). This tendency is understandable, of course, in the light of psychologists' interest in studying human behavior in simplified laboratory situations. A second line of study has not involved subjective-probability models as such, but only pieces of models, such as subjective linear predictors. (See the first half of Slovic and Lichtenstein 1971, or Dawes 1971). This would seem to arise from psychologists' primary interest in descriptive theories of overt behavior, rather than in helping experts to express their opinions more usefully. Psychologists are also not generally familiar with the mathematics of parameterized or multivariate probability *distributions*, although they may be conversant with high dimensional linear structures and data analyses. Garthwaite (1977) gives a review of the psychological literature.

Finally, I cannot resist commenting on Professor Novick's discussion of Professor Good's paper, in which he writes, "the problem of assessing a reasonable prior distribution for the final stage in a hierarchical model is a difficult one". In fact, the problem can be solved immediately in principle by the methods of Kadane, Dickey, Winkler, Smith and Peters (1978). I hope to write a short paper soon on this, but in case I do not get around to it, I shall outline the main idea here.

Although presented in the guise of the problem of assessing a Bayesian predictive distribution for the normal regression sampling model, the methods of Kadane *et al* (1978) are fully general for the assessment of a distribution in any given multivariate location-scale family. The predictive distribution for a hierarchical normal linear model with proportional unkown variance components is multivariate *t*, but with special constrained form of location vector and scale matrix. Need I say more? Since one can assess a fully general multivariate *t* distribution, one can surely then take the further step of approximating it mathematically by the nearest distribution (in a suitable norm) of the hierarchical predictive form. Also, the differences between the two distributions can help indicate whether the hierarchical structure is a realistic model assumption.

I.J. GOOD (*Virginia Polytechnic Institute and State University*):

I would like to thank the discussants both for their generous comments and for their critiques which force me to clarify a few points.

I had hoped that my views about kinds of probability were well enough known so as not to have to repeat much of them in my article (see, for example, Good 1965, pages 6 and 7). Dr. Lindley seems to agree with me that it is at least convenient to talk about "chances" (physical probabilities) in addition to subjective (personal) probabilities, whether you believe in the independent "existence" of physical probabilities or whether you arrive at their "as if" existence by using de Finetti's theorem. All I have further claimed is that some people (such as H. Jeffreys, Carnap and Jaynes) find it helpful to assume the existence of credibilities or logical probabilities (unique degrees of belief of a perfectly rational being). I agree with Drs. Lindley and DeGroot that subjective probabilities are the most fundamental and, to quote the editors of *Science* in their heading for Good (1959), "Although there are at least five kinds of probability, we can get along with just one kind". So there doesn't seem to be very much basic philosophical difference between my position and Dr. Lindley's. By a "subjective probability" I mean one belonging to a body of beliefs which you have made some attempt to make reasonably coherent. When no such attempt is made I refer to "psychological probability" There is a continuous gradation from psychological probabilities to coherent subjective probabilities, the latter being something of an ideal though a little less so than logical probabilities or credibilities. Whether you prefer to adopt one ideal or the other or both is a matter of taste. If you wish, you can replace "logical probabilities" in my exposition by "your true coherent subjective probabilities that are 'struggling to get out'" This substitution has no affect at all on the hierarchical technique, which is intended to be a psychological aid. Surely, we all aim to have "objective judgments" in some sense.

At a slightly less fundamentally philosophical level my position is perhaps different from those of Drs. Lindley and DeGroot (and from de Finetti) in my support of a Bayes/non-Bayes compromise or synthesis. I have heard Dr. Lindley imply, at least as far back as 1967, that there is no reason to expect any relationship between tail-area probabilities and Bayes factors. I do expect such an approximate relationship *in those circumstances where the Bayesian model and the "tail-area" (or "Fisherian") model are both sensible*; the reason being, in my opinion, that the sensible tail-area merchant is implicitly approximately Bayesian at the back of his mind. He doesn't usually consider null hypotheses that are highly improbable, and, when he does, he insists on a smaller tail-area probability. And he does so insist also, if he is sensible, when the losses due to errors dictate such a policy, and also if his sample is extremely large. And he selects single-tail or double-tail probabilities depending on some concept of the non-null hypothesis. In other words, if he is sensible, he is something of a Bayesian without realizing it or without admitting it, or, if he is me, *while* admitting it. As I have said on several occasions, "To the Bayesian all things are Bayesian". For a further discussion of the relationship betweem tail-area probabilities and Bayes factors see Good (1980a).

Dr. DeGroot asks in what sense the higher-level probabilities are more woolly. I mean that they are more controversial and can be judged less sharply as measured, for example, by the ratio of the upper to the lower odds (where "odds" = $p/(1\text{-}p)$, p being a probability). My assertion in Good (1952) that "the higher the type the less the woolliness matters" was based on the belief that agreement is often reached among scientists, all of whom are Bayesians at heart without necessarily knowing it.

The point of a Bayes/non-Bayes compromise is to obtain greater agreement between the polarized camps, not to question the essential truth and inevitability of the neo-Bayesian philosophy of partially-ordered subjective probabilities. (What I say fifty times is true.) Likewise the unity that this philosophy gives to statistics, and to the philosophy of science, is not undermined by talking about three kinds of probability. Unitarians and trinitarians can live in peaceful coexistence and do not need to indulge in holy wars.

Dr. DeGroot points out that in hierarchical models involving physical probabilities we are interested in the higher-level probabilities as well as those at the lowest level. This is to some extent true also for subjective models, for example, in the multinomial (not histogram) work the marginal distribution of the flattening constant k is of interest, especially its (Type II) Maximum Likelihood value. But approximate values are adequate.

Turning now to Dr. Leonard's informative comments in which, for about the first time in history, a skittles game of chess is partly immortalized in print, I must begin by admitting an ambiguity in my use of the term "rough", and I can see that this ambiguity has led to a misunderstanding. When successive or close observations are initially likely to be highly correlated, as in a continuous density curve or histrogram, perhaps the term opposite to smooth (or "ordinally smooth") should be "ordinally rough" or "flamboyant" (in its architectural meaning). The roughness penalty in the work on probability density estimation refers to this ordinal roughness. But in my work on multinomial distributions and contingency tables I have used the term "rough" to imply a deviation from equiprobability; for example, multinomial probabilities $p_1,p_2,$

more accurate than those given by Good, 1967, p. 415 where there was one misprint.)

k_{max}	1.05	.09	.88	3.35	18.8	12.9	140
k_0	1.2	.24	1.0	4.6	20	14.5	59
k_0'	1.7	.30	.77	3.9	18.4	11.3	236
k_{TL}	1.7	.20	.76	3.8	18.3	11.2	236

k_{max}	.044	16.3	6.1	3.3	.39	.64	.15
k_0	.08	17	7.0	3.6	.77	1.1	.24
k_0'	.08	16.4	6.2	3.3	.90	.62	.35
k_{TL}	.07	15.9	5.7	2.8	.80	.52	.25

The approximations k_0' and k_{TL} are negligibly different and it is not clear whether they are better at estimating k_{max} or k_0. The approximations are better than they look when used for approximating G. For instance, in the first example $G = 3.96$, while taking k_{max} as 1.7 instead of 1.05 gives the approximation $G \approx 3.80$. Likewise when k_{max} is large, the null hypothesis $k = \infty$ cannot be rejected, so, in the seventh example, the estimates k_0' and k_{TL} are far better than they look. One needed to do the theory properly to evaluate the approximations so I cannot agree with Dr. Leonard's comment that the approximation "avoids" much of the theory. It might avoid it *in the future* but the theory had to be done to justify the approximations.

Provoked by the goodness of the approximation and by the consideration of Stone (1974) and Leonard (1977), I wondered whether k_{max} for contingency tables could be approximated by $u(N\text{-}1)(u+1)^{-1}(X^2\text{-}u)^{-1}$ (or by ∞ when this is negative). [Here $u = (r\text{-}1)(s\text{-}1)$.] Calculations by Dr. J.F.Crook indicate that it leads to good approximations to G as computed using the full theory of Good (1976) and of Crook and Good (1980). The approximations are about as good as for the multinomial problem. For more details see Good (1980b). It would be interesting to know whether the results would be even better if we replaced our mixture of Dirichlets, given $\bar{H}$, by the corresponding mixture of densities proportional to $\Pi p_{ij}{}^{k_{ij}-1}$, where k_{ij} is of the form $\alpha n_{i.} n_{.j} N^{-2}$. Moreover the approximation $u(N\text{-}1)(u+1)^{-1}(X^2\text{-}u)^{-1}$ for k_{max} might be good for multidimensional contingency tables, where u is the number of degrees of freedom.

When Dr. Leonard says that F might be sensitive to the choice of hyperprior [if ridiculous hyperpriors are used?] he is saying something with which I have long been familiar (Good, 1967, p. 405). Moreover it does not depend on going outside the log-Cauchy family. As I said, no improper hyperprior can be used, because $F(k) \to 1$ as $k \to \infty$, and therefore of course proper priors that are close enough to being improper will give a resultant Bayes factor F arbitrarily close to 1. If this leads anyone to abandon the hierarchical Bayes methodology, then he might as well abandon Bayesianism itself (because some ridiculous priors kill any given evidence), and therefore, if he is consistent, he should abandon statistics, science and rationality.

I was pleased to receive Dr. Geisser's additional historical references which I had overlooked. I agree entirely with his belief in the value of the method of predictive sample reuse (= cross-validation). Regarding his second comment, he is certainly right that one can sample a contingency table in more than three ways. But, from a Bayesian point of view, the question of significance of departure from independence does not depend on whether, for example, the row totals are sampled as a negative binomial or as an ordinary binomial, because optional stopping does not affect the evidence once the sampling is done. The way I have often expressed this since 1950, in conversations with other statisticians, is that the evidence cannot depend on whether the statistician is telling the truth when he says he stopped sampling because he had to catch a train. If a Bayesian were to check whether there was really a train due, at the time the other statistician claimed, it would not be because this would affect the evidence directly. Rather it would be to obtain evidence of whether the other statistician was an honest man. If the statistician *thinks* he is cheating when he pretends he has a train to catch, then he can be convicted of *attempted* cheating. A statistician who tries to cheat in one respect may very well have succeeded in cheating in another respect.

Dr. Novick's general agreement with my views is encouraging. He raises the matter of the relationship between Bayesian hierarchical methods and empirical Bayes methods. I had described what I understand by the empirical Bayes method: one interpretation of it is the use of objective physical empirical frequency-based priors, as by von Mises (1942), and another somewhat different interpretation is the one favored by Krutchkoff and exemplified by Turing's idea of 1941 which I embellished and published in 1953. On the other hand, in the hierarchical Bayes methods, as I understand them, the use of the hierarchy is a psychological aid to a good Bayesian judgement of subjectivistic priors. If the hyperparameters (or hyperhyperparameters) are estimated by means of a non-Bayesian method, such as by Type II Maximum Likelihood, then one has a Bayes/non-Bayes compromise or pseudo-Bayesian procedure which is not empirical Bayesian by either of the two definitions just mentioned. The essential point, as I see it, is that the empirical Bayes methods are not supposed to involve epistemological probabilities, but only physical probabilities. They are Bayesian only in the sense of using Bayes's theorem and I think Dr. Novick would agree with this usage. I endorse his wish that people would not call methods empirical Bayes when they are pseudo-Bayesian. I agree too that pseudo-Bayesian methods, as so far used, usually provide point estimates of parameters and variables, but it seems to me that they can be extended in natural ways to provide interval estimates. For an interval estimate, obtained non-Bayesianly, at the top of a hierarchy, will lead to interval estimates lower down.

Dr. Novick said it would be nice to have had a discussion of the relationship between kinds of probability. Certainly the topic is closely related to the topic of my paper, and it is one of my interests, but, for the sake of brevity, I did not discuss it. There is a little discussion of it in, for example, Good (1962b, 1965, 1975). For discussion and classifications of kinds of probability see Good (1950, 1952, 1959, 1965, 1966, 1971d, and 1975b), where citations of earlier writers are given, including Carnap, Poisson, and Kemble.

REFERENCES IN THE DISCUSSION

DAWES, R.M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *Amer. Psychol.* **25**, 180-188.

GARTHWAITHE, P. (1977). *Psychological aspects of subjective probability elicitation.* M.Sc. Thesis. Department of Statistics, University College of Wales, Aberystwyth.

GEISSER, S. (1975a). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70**, 320-328.

— (1975b). A new approach to the fundamental problem of applied statistics. *Sankhya B* **37**, 385-397.

GOEL, P.K. and DEGROOT, M.H. (1979). Information about hyperparameters in hierarchical models. *Tech. Rep.* **160**. Department of Statistics, Carnegie-Mellon University.

GOKHALE and PRESS, S.J. (1979). The assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *Tech. Rep.* **58**, University of California, Riverside.

GOOD, I.J. (1962). A compromise between credibility and subjective probability. *International Congress of Mathematicians, Abstracts of Short Communication.* Stockholm, 160.

— (1965b). Speculations concerning the first ultraintelligent machine. *Advances in Computers* **6**, 31-38.

— (1971d). Unpublished lecture notes entitled "The Bayesian Influence" 122. Statistics Department, Virginia Polytechnic Institute and State University.

— (1972). Food for thought. In *Interdisciplinary Investigation of the Brain* (J.P. Nicholson, ed.) 1972, 213-228. New York: Plenum Press.

— (1975b). Explicativity, corroboration and the relative odds of hypotheses. *Synthese* **30**, 39-73.

— (1980a). The logic of hypothesis testing. In *Philosophical Foundations of Economics* (J.C. Pitt, ed.) Dordrecht: Reidel.

— (1980b). An approximation of value in the Bayesian analysis of contingency tables. *J. Statist. Comput. Simulation* (in press).

KADANE, J.B., DICKEY, J.M., WINKLER, R.L., SMITH, W.S. and PETERS, S.C. (1980). Interactive elicitation of opinion for a normal linear model. *J. Amer. Statist. Assoc.* **75** (to appear).

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60**, 297-308.

— (1975). Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. B* **37**,23-37.

— (1977). A Bayesian approach to some multinomial estimation and pretesting problems. *J. Amer. Statist. Assoc.* **72**, 869-874.

LINDQVIST, B. (1977). How fast does a Markov chain forget the initial state? A decision theoretical approach. *Scand. J. Statist.* **4**, 145-152.

— (1978). On the loss of information incurred by lumping states of a Markov chain. *Scand. J. Statist.* **5**, 92-98.

O'HAGAN, A and LEONARD, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* **63**, 201-203.

SCOTT, D., TAPIA, R.A. and THOMPSON, J.R. (1978). Multivariate density estimation by discrete maximum penalized likelihood methods. In *Graphical Representation of Multivariate Data.* 169-182. New York: Academic Press.

SLOVIC, P. and LICHTENSTEIN, S.C. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior and Human Performance* **6**, 649-744.

STONE, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61**, 509-515.

TVERSKY, A. (1974). Assessing uncertainty. *J. Roy. Statist. Soc. B* **36**, 148-159.

Trab. Estadist. (*1980*)
Vol. **31**, pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (*1980*)
pp. 521-534 Valencia, Spain: University Press.

12. Bayesian non-parametric theory

INVITED PAPER

DALAL, S.R. (*Rutgers University*)
Non-parametric Bayes Decision Theory

DISCUSSANTS

GOOD, I.J. (*Virginia Polytechnic Institute and State University*)
KADANE, J.B. (*Carnegie-Mellon University*)
LEONARD, T. (*University of Warwick*)
O'HAGAN, A. (*University of Warwick*)
SMITH, A.F.M. (*University of Nottingham*)

REPLY TO THE DISCUSSION

Nonparametric Bayes Decision Theory

S.R. DALAL
Rutgers University[*]

SUMMARY

A summary of the seminar with the same title is presented. Ferguson's fundamental work on the theory of Dirichlet processes is elucidated and their shortcomings are discussed. Some modifications are also proposed and illustrated. Some of the intricate mathematical issues related to the definitions and the proofs are not discussed for the sake of clarity and brevity. The development related to unimodal processes, briefly mentioned in the last section, will appear as a joint work with Professor W.J. Hall elsewhere.

Keywords: NONPARAMETRIC BAYES; DIRICHLET PROCESSES; DIRICHLET INVARIANT PROCESSES; ESTIMATION OF A SYMMETRIC C.D.F.; ESTIMATION OF A UNIMODAL DENSITY; ESTIMATION OF A POINT OF SYMMETRY.

1. INTRODUCTION

Nonparametric theory deals with the problems of inference when the underlying distribution is not specified in terms of a parametric family. This theory can be gainfully employed in many situations as models are seldom more than approximations to reality, and the procedures which are optimal for a given parametric family (i.e. the 'Idealized Model') may perform poorly even for models which are near to the idealized model (e.g. Tukey (1960), Huber (1964)).

However, 'classical' nonparametric theory disregards much of the existing knowledge about the idealized model. Further, evaluations and comparisons are usually carried out asymptotically at specific parametric models.

To avoid these shortcomings, it is useful to think that there is an idealized model and that the observed distribution is a (possibly random) perturbation of the idealized model. This approach has been used by Huber (1972) and

[*] Presently at Bell Labs, Murray Hill, N.J. 07974, U.S.A.

others to create an elegant theory of robustness. Here, we explore an alternative approach of Ferguson (1973), who derived, and suggested the use of Dirichlet processes as priors for nonparametric problems. Specifically, we shall review Dirichlet processes (Section 2), note their anomalies and inadequacies (Section 3), and suggest some modifications.

2. DIRICHLET PROCESSES

Let $(\chi, B(\chi))$ be the sample space and P^* be the space of all probability measures on $(\chi, B(\chi))$. P^*, the parameter space for many nonparametric problems is quite large and consequently many procedures turn out to be minimax. Hence, the Bayes criterion becomes more relevant.

For the Bayes framework, it is necessary to consider a class of priors over P^*, i.e., a class of *random probability measures,* which is a) *mathematically tractable,* b) *rich*, and c) *easy to parameterize.* Several procedures have been suggested toward this end, notably by Dubins and Freedman (1966), Kraft and Van Eeden (1964), Rolph (1968), Ferguson (1973), Doksum (1974) and Sethuraman (1979). In statistical inference Ferguson's priors, Dirichlet processes, have been more often used that the other procedures, because of their intuitive properties and tractability, e.g., Ferguson (1974), Susarla and Van Ryzin (1976), Phadia and Susarla (1979), Berry and Christensen (1979).

Mixtures of Dirichlet processes, proposed by Antoniak (1974), have also been used in Bio-assay and regression-type problems. Relatively few applications not related to Dirichlet processes are available. For example, Ferguson and Phadia (1979) have dealt with censored data problems using Doksum's *neutral to right processes.* Also, some new non-Dirichlet-process priors developed by Sethuraman (1979) may prove to be useful. We shall, however, follow Ferguson's approach with some modification. Before delineating the modifications, we define and briefly state some elementary properties of Dirichlet processes below.

Definition. A random probability P is a Dirichlet process if there exists a finite, finitely, additive measure, α, such that for every measurable partition $B_1, \ldots, B_k$ of χ, $(P(B_1), \ldots, P(B_k))$ has a Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$. We then write $P \epsilon DP(\alpha)$ and denote the corresponding random probability measure by P.

Elementary properties. Let $\alpha = M \cdot Q$, where M is a positive number, and Q is a probability measure on $(\chi, B(\chi))$. Then $P \in D(\alpha)$ implies that $\mathcal{E}P = Q$. Q, thus, can be thought of as ideal distribution. Further, the number M can be viewed as the prior example size (e.g. Novick and Hall, 1965). Using these properties DP priors can be *easily parameterized.*

The second desirable property, Richness, mentioned earlier, has two aspects. First, richness of support is essential to deal with a broad class of nonparametric problems. Secondly, if one is to restrict attention to a specified class of priors, it is essential for this class to have members capable of approximating any prior belief. We call this latter aspect *adequacy*. Both of these issues can be examined by imposing a 'natural' topology on P^*, the space of all probability measures, and P^{**}, the space of all random probability measures.

For a lack of the "natural" topology, various topologies can be considered (e.g. Ferguson (1973), Dalal (1978), Dalal and Hall (1980)). By considering the weak* topology on P^* obtain by imbedding P^* on the product space $[0,1]_{\beta}^{(\chi)}$, Ferguson (1973) showed that all α-absolutely continuous distributions are in the support of $DP(\alpha)$. Dalal (1978) showed that this kind of imbedding leads to random probability measures which select finitely additive probability measures on $(\chi, B(\chi))$ with probability one. Further, although the class of Dirichlet processes is not *adequate* in terms of approximating a given belief, a convex hull of this class of mixtures of Dirichlet processes (MDP) is adequate in this regard (see Dalal (1978), Dalal and Hall (1980)).

The mathematical tractability of any class of priors can only be ascertained by examining the ease with which the posterior and various simple expectations are obtained. With respect to Dirichlet processes, Ferguson showed that, given a sample $X_1, \ldots, X_n$, the posterior is $DP(\alpha + \Sigma\delta_{x_i})$, where δ_x is the unit mass degenerate at x. This conjugate prior property has been extensively used in applications.

3. SHORTCOMINGS AND MODIFICATIONS WITH APPLICATIONS

First, we discuss an anomaly (discreteness), and an inadequacy (to deal with invariant problems) of Dirichlet processes. This is followed by some modifications to overcome these defects. A few illustrative example are also given.

3.1. *Shortcomings*

i) **Discreteness.** It is known that DP's are discrete with probability one (e.g. Blackwell (1973), Berk and Savage (1977)). This discreteness is more than a technical aberration. In some applications this has led to non-intuitive answers. Further, the posterior changes the masses only at the observed sample points. Intuitively, however, it would be appealing to have a prior which increases the probability of a neighborhood instead.

ii) **Invariance.** In nonparametric problems, one is permitted to have nonparametric beliefs, e.g. symmetry of the underlying distribution (i.e. in the one-sample problem), exchangeability, spherical symmetry, etc. However *DP* and the other priors defined so far live on the class of all probability measures. It would be useful to also have priors giving probability one to invariant (under a group *g*) probability measures.

3.2. *Modifications*

Our approach, simply stated, is to modify the sample paths of a *DP* (i.e. the distributions selected by a *DP*) to eliminate these shortcomings. The modified process, although closely related to the *DP*, is usually more complex. However, one can use the updated version of the *DP* to manipulate the posterior of the modified process.

3.2.1. *Modifications related to Invariance*

Let $g = \{g_1,\dots,g_k\}$ be a finite group of measurable transformations on $(\chi, B(\chi))$ and P be a random probability measure. Define a new random probability measure Q by the mapping $Q(A) = \frac{1}{k}\Sigma P(g_iA)$. Clearly Q selects g-invariant distributions with probability one. Such a scheme can also be used with a compact topological group to obtain invariant distributions with probability one. When P is a Dirichlet process with g-invariant α, Q is called the Dirichlet Invariant process ($DIP(\alpha)$). These kind of processes have been studied in Dalal (1979a). The behavior of *DIP*s is similar to *DP*s, e.g. if $x_1,\dots,x_n$ is a sample from $Q \epsilon DIP$, then $Q|X_1,\dots,X_n$, the posterior of Q, is $DIP(\alpha + k^{-1}\Sigma\Sigma g_iX_j)$.

Using *DIP*'s Bayes estimates of various functionals can be obtained. Some illustrative applications are considered below.

i) **Estimation of a symmetric c.d.f.** Consider the problem of estimating a c.d.f., F_μ symmetric about a known point μ. Let the loss function be $L(F_\mu,\hat{F}) = \int (F_\mu(t) - \hat{F}(t))^2 dW(t)$, where W is a finite prespecified weight function. For Bayes estimation, consider the prior $DIP(\alpha)$, where $\alpha = M \cdot Q$ and the group g is $\{g_1, g_2\}$; $g_1(x) = 2\mu - x$, $g_2(x) = x$. Let G be the c.d.f. corresponding to Q. The Bayes estimate then can be shown to be a convex linear combination of the initial guess G and the μ-symmetrized empirical c.d.f. $\hat{F}_n$ (Dalal, 1979a), i.e.

$$\hat{F}_\mu = \frac{M}{M+n} G + \frac{n}{M+n} \hat{F}_n.$$

As n becomes larger, the dependence on the initial guess G becomes weaker. Also the expression for $\hat{F}_\mu$ suggests that M can be thought of as the prior sample size, as discussed earlier.

The above Bayes formulation can also be exploited to get a minimax estimate,

$$\hat{F}_\mu = \frac{1}{4(1+\sqrt{n})} + \frac{1}{2(\sqrt{n}+n)} \Sigma(\delta_{x_i} + \delta_{2\mu - x_i}) + \frac{1}{2(1+\sqrt{n})} \delta_\mu .$$

Bayes estimates of F_μ for μ unknown have also been obtained in Dalal (1979a).

ii) **Estimation of the unknown center of symmetry.** Consider the usual one sample problem of estimating the center of symmetry of an arbitrary distribution F. Specifically, assume the following model $Y_i = \mu + \Delta_i$ where the Δ_i are i.i.d. with an arbitrary distribution, F, symmetric about 0. For the Bayes formulation, Let $F \epsilon DIP(\alpha)$, $\alpha = M \cdot G$, and μ have the non-informative uniform distribution over the reals. Then the Bayes estimate using squared error loss $(\mu - \hat{\mu})^2$ can be found. In the case of the idealized model, G, being standard normal (density φ), and assuming all distinct observations the Bayes estimate $\hat{\mu}$ is (Dalal, 1979b)

$$\hat{\mu} = \frac{M}{M+n} \bar{y} + \frac{n}{M+n} \mu^* ,$$

where

$$\mu^* = (\sum_{i<j} w_{ij} \frac{y_i + y_j}{2} / \sum_{i<j} w_{ij})$$

and

$$w_{ij} = (\prod_{k \neq i,j} \varphi(y_k - \frac{y_i + y_j}{2}))(\varphi \frac{y_i - y_j}{2}).$$

μ^* is a weighted mean of pairwise averages. The weight given to the pair $\frac{y_i + y_j}{2}$ is inversely proportional to the distance between y_i and y_j, and the distance of $\frac{y_i + y_j}{2}$ from the rest of observations. Thus, this estimate is robust. Numerical and other investigations on this estimate are considered in Dalal (1979b).

3.2.2. *Modifications related to continuity unimodality*

In density estimation problems, the usual estimates are obtained by smoothing out the functionals of the empirical c.d.f. This is usually accomplished by convoluting with different kernels.

This kind of idea can be used to smooth out the sample paths of the *DP*'s followed by sampling from the smoothed process. Some of these ideas have been considered in a joint work with Professor W.J. Hall of the University of Rochester. Using this, Bayes estimates of general densities, unimodal densities, modes, etc., can be obtained. Some details have been worked out in Dalal and Hall (1977).

ACKNOWLEDGEMENTS

I wish to thank Professor W.J. Hall for various helpful discussions. This research was supported by NSF Grant MCS 78-02160 and Rutgers University Research Council.

REFERENCES

ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Ann. Statist.* **2**. 1152-1174.

BERK, R.H. and SAVAGE, I.R. (1977). Dirichlet processes produce discrete distributions: An elementary proof. *Tech. Rep.* Rutger University.

BERRY, D.A. and CHRISTENSEN R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7**, 558-69.

BLACKWELL, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1** 356-358.

DALAL, S.R. (1978). A note on the adequacy of mixtures of Dirichlet processes. *Sankhya, A*, **40**, 185-91.

— (1979a). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stoch. Proc. and Appl.* **9**, 99-107.

— (1979b). Nonparametric and robust Bayes estimation of location. *Proc. Optimizing Methods in Statistics.* 141-166. New York: Academic Press.

DALAL, S.R. and HALL, G.J., Jr. (1980). On approximating parametric Bayes models by nonparametric Bayes models. *Ann. Statist.* **8**, 664-672.

DALAL, S.R. and HALL, W.J. (1977). Unimodal density estimation. *Tech. Rep.* Rutgers University.

DOKSUM, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Prob.* **2**, 183-201.

DUBINS, L.E. and FREEDMAN, D.A. (1966). Random distribution function. *Proc. 5th Berkeley Symp.* **2**. 183-214. Berkeley: University of California.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

— (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.

FERGUSON, T.S. and PHADIA, E.G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7**, 163-86.

HUBER, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**; 73-101.

— (1972). Robust statistics: a review. *Ann. Math. Statist.* **43**, 1041-1067.

KRAFT, C.H. and VAN EEDEN, C. (1964). Bayesian bio-assay. *Ann. Math. Statist.* **35**, 886-890.

NOVICK, M.R. and HALL, W.J. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**, 1104-17.

PHADIA, E.G. and SUSARLA, V. (1979). An empirical Bayes approach to two sample problem with censored data. *Comm. in Statist.* **8**, 1327-1351.

ROLPH, J.E. (1968). Bayesian estimation and mixing distributions. *Ann. Math. Statist.* **39**, 1289-1302.

SETHURAMAN, J. (1979). Personal Communication.

SUSARLA, V. and PHADIA, E.G. (1976). Empirical Bayes Testing of a distribution function with Dirichlet process priors. *Comm. in Statist. A*, **5**, 4505-69.

SUSARLA, V. and VAN RYZIN, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897-902.

TUKEY, J.W. (1960). A Survey of Sampling from Contamined Distributions. In *Contributions to Probability and Statistics.* (Olkin ed.) Stanford: University Press.

DISCUSSION

I.J. GOOD (*Virginia Polytechnic Institute and State University*):

Dr. Dalal discussed the application of the Dirichlet-process priors to continuous problems. In my work on categorical data, I found it necessary to use *mixtures* of Dirichlet priors (Good, 1965, 1967, 1976; Good & Crook, 1974; Crook & Good, 1980). In Good (1978) I asked whether it would be useful to use mixtures of Dirichlet processes for continuous data, such as for testing independence in continuous bivariate distributions. Also, can we apply "Ockham's hyperrazor" by somehow selecting the Dirichlet processes so that only one hyperparameter is required? If so, this could be given a hyperprior as in the categorical work.

J.B. KADANE (*Carnegie-Mellon University*):

One of the interesting things in non-parametric statistics is the interpretation of various interesting quantities as *U*-statistics. For example, Wilcoxon's statistic is an estimate of $P[X<Y]$. Have the modifications of Dirichlet processes been studied to see whether Wilcoxon's statistic can be justified as an estimator from this point of view?

T. LEONARD (*University of Warwick*):

Professor Dalal's convolution of the Dirichlet process seems to me to involve some really brilliant ideas. It will be regarded as one at the important contributions in the area of non-parametric density estimation. His generalisation of the Dirichlet process avoids the pitfalls faced by Ferguson, for example the problems of spiky posterior estimates and specific prior covariance structures. His prior distribution is very general and simply formulated and leads to appealingly smooth posterior estimates. He is to be congratulated on achieving an original idea of such beautiful and wide-ranging simplicity.

When specifying his prior distribution, I think that it would be helpful if Professor Dalal worked in terms of his prior covariance kernel, as well as his prior mean value function, since this would demonstrate precisely how he intends to smooth his estimates. This would also highlight the similarity between his approach and that of the early work of Whittle, who just specified the first two-moments of his prior. By completely specifying his prior Dalal achieves the same generality as Whittle, but he does not run into problems of negative posterior estimates, and he is also able to make posterior probability statements about the unknown density, as well as providing point estimates.

Professor Dalal's posterior estimates are constrained to the class of kernel estimates and I wonder whether this is a property of the type of prior distribution assumed? My own approach constructs a prior in logit space where it seems very natural to think in terms of linear relationships and covariance kernels, and my estimates assume a general non-linear form rather different from kernel estimates. The following rather undesirable properties of kernel estimates are avoided under my approach:

1) The overspread-out nature of kernel estimates (the estimated variance is always greater than the sample variance)
2) The dependence of bandwidth upon sample size in order to achieve asymptotic consistency, or under Whittle's approach the contraction to delta functions as the sample size increases.
3) The problem that when there are only a moderate number of observations kernel estimates will either oversmooth or possess bumps in the tails.

I think that the great strenght of a Bayesian approach to nonparametric density estimation lies in the fact that it permits us to model the density via its prior estimate whilst avoiding any constraint on the posterior estimate to belong to a parameterized family. It for example provides a particularly viable alternative to classical tests for fit, since we simply need to investigate differences between the posterior estimate and the prior hypothesised estimate.

A. O'HAGAN (*University of Warwick*):

Professor Dalal has shown us a very interesting formulation of nonparametric inference. The so-called nonparametric problems are characterised in his approach, and in the earlier work of Ferguson, by a vast number of parameters. I believe this feature is inevitable: even when inference centres on some subparameter like the median, Professor Dawid has shown in his paper at this meeting that nuisance parameters cannot be dismissed without careful consideration. Given that there really are infinitely many parameters, *only* a Bayesian approach is feasible. The problem is underidentified (or overparametrised) and no amount of data will give sufficient information to render the prior irrelevant. In particular, the way in which the prior relates parameters to each other influences strongly the shape of posterior inferences. Ferguson's Dirichlet process, for example, yields discontinuous posterior means. It is not enough that the prior should look sensible; it must also give sensible posterior inferences, and it is quite proper for Professor Dalal to seek for priors which give posterior inferences having sensible shapes.

A.F.M. SMITH (*University of Nottingham*):

I hope that all who have enjoyed Professor Dalal's elegant presentation and admired his undoubted mathematical ingenuity will forgive me if I express the philistine sentiment that exercises involving contemplation of completions of spaces of mixtures of Dirichlet processes have very little to do with interpreting data in the light of personal judgment, and, whatever else they are, are *not* Bayesian Statistics.

Instead of seeking a tractable way of representing the uncontemplatable (i.e. measures having large support over gigantic spaces of distributions), we should first of all decide what aspects of the problem we *are* able to contemplate and *then* seek a tractable representation.

As an example of what I have in mind, suppose we want to make inferences about location, given up to 50 observations from an (unknown) member of the (assumed) location-scale family. I *can* contemplate qualitative features that may be relevant -such as heaviness of tails, skewness, etc.- and I *can* realize that with samples of this size there is little point in seeking a prior measure with large support in the location-scale family. (We simply cannot distinguish other than quite crude qualitative differences between distributions.) Instead, a sufficiently rich mixture should result from a prior with a sensibly chosen representative *finite* support. One such crude choice which incorporates heavy, and light-tailed departures from Normality, together with skewness in both directions, is a finite mixture model consisting of the Normal, Uniform, Laplace, Right-Exponential and Left-Exponential distributions.

This has the added advantage that all the necessary Bayesian manipulations can be carried out analytically. (See Spiegelhalter, 1978.)

I have always understood "Nonparametric" to mean "Enormous Parameter (Model) Space", where "enormous" signifies "too big to have to think meaningfully about". I suggest, therefore, that we should be very circumspect about any theory which couples "Nonparametric" with the word "Bayesian".

S.R. DALAL (*Rutgers University*):

Professor Good during his discussion at the conference inquired about the suitability of symmetric Dirichlet distributions and associated processes as priors for nonparametric problems. Use of these priors in contingency tables leads to manageable numbers of hyperparameters and some ease in numerical computations (Good, 1976). Unfortunately, in many interesting nonparametric problems, the interesting sets are of various sizes, and thus, the kind of symmetry inherent in contingency tables is absent. This rules out the use of symmetric Dirichlet distributions. However, as indicated in the paper we can use Dirichlet symmetric processes whenever some appropriate invariance structures can be assumed. Professor Good's comment on the use of "Ockham's hyperrazor" needs further investigation.

Professor Kadane has raised an interesting and an important issue related to justification of classical nonparametric procedures based on U statistics through the nonparametric Bayes theory. This line of inquiry has already been followed in Professor Ferguson's fundamental paper. He showed that in the problem of estimation of $\int FdG$ with a squared error loss, the Bayes estimate is a convex linear combination involving the Mann-Whitney statistic. Similar justification can be provided for several other nonparametric procedures. For example, my work with Professor Phadia has shown that Kendall's π can also be similarly interpreted from Bayesian point of view.

Dr. Leonard has been very kind in praising my work on density estimation. The applicability and usefulness of my approach can be judged only after examining the complexity of the estimates, the large sample properties (e.g. consistency rates of convergence), etc. In this regard, the references furnished by Dr. Leonard to his work (1973), Whittle's work (1958) and Good and Gaskin's work (1971) will be very useful.

Dr. Leonard is also quite correct in pointing out that the posterior estimates are constrained to the class of kernel estimates because of the nature of the prior. However, in the important problem of unimodal density estimation this is not a constraint. Dr. Leonard has also been able to convince me that it would be helpful to work in terms of covariance kernels. I think this deserves detailed investigation.

I do concur with Dr. O'Hagan's comment on the inevitability of the parametrization by large number of parameters in Bayes formulation of nonparametric problems. This is not to say that in such a formulation no amount of data will give sufficient information to render the prior irrelevant. In fact, I think that some sort of generalized version of the theory of precise measurement would hold and accordingly the precise nature of the large number of parameters involved would be unimportant.

Professor Smith comments that we would be circumspect about any theory which couples 'Nonparametric' with the word 'Bayesian'. I disagree with his logic. Much recent works shows that suchs an alliance is not an unholy one. This is also best illustrated in the usual one sample problem where observations are obtained as differences of pairs of measurements. Here the assumption of symmetry is easily justified and beliefs about the point of symmetry may also be easily parametrized. Savage's theory of precise measurement tells us that the precise formulation of beliefs about the point of symmetry is immaterial. However, an incorrect specification of the model does have serious consequences for the Bayesian (e.g. Berk, 1966). In this instance, whithout any additional information, the Bayesian nonparametric theory is certainly a viable contender to any other form of Bayes analysis. Also, if Dirichlet symmetric processes are used as priors, then a generalization of Savage's theory of precise measurement suggests that the parameter α of the process need not be precisely specified.

Professor Smith also contends that the results related to completion of spaces of mixtures of Dirichlet processes are not part of Bayesian statistics. This may be true in a narrow sense. However, disregarding its Bayesian implications will be a mistake. The result which Professor Smith refers to says that a Bayesian, in quest for a suitable prior for a nonparametric problem, need not go beyond the class of mixtures of Dirichlet processes. A parametric counterpart would say that the Bayesian need not go beyond the class of mixtures of natural conjugate priors. (Dalal and Hall, 1977).

REFERENCES IN THE DISCUSSION

BERK, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37**, 51-58.

CROOK, J.F. and GOOD, I.J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part II. *Ann. Statist.* (in press).

GOOD, I.J. (1965). *The estimation of Probabilities: An Essay on Modern Bayesian Methods.* Cambridge, Massachusetts: The M.I.T. Press.

— (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. B* **29**, 399-431 (with discussion). *Corrigendum* **36** (1974), 109.

— (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**, 1159-1189.

— (1978). Review of Ferguson, Thomas S., "Prior distributions on spaces of probability measures". *Ann. Statist.* **2**, (1974) 615-629; *Math. Rev.* **55**, 1546-1547.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**, 711-720.

GOOD, I.J. and GASKINS, R.A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255-277.

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60**, 297-308.

SPIEGELHALTER, D.J. (1978). *Adaptive inference using a finite mixture model.* Ph.D. Thesis. London: University College.

WHITTLE, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. B* **20**, 334-343.

Trab. Estadist. (*1980*)
Vol. **31** pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M., eds.
Bayesian Statistics (*1980*)
pp. 535-582 Valencia, Spain: University Press

13. Coherence of models and utilities

INVITED PAPERS

LEONARD, T. (*University of Warwick*)
The roles of inductive modelling and coherence in Bayesian statistics

NOVICK, M.R., DEKEYREL, D.F. and CHUANG, D.T. (The University of Iowa)
Local and regional coherence utility assessment procedures

DISCUSSANTS

DICKEY, J.M. (*University College Wales*)
DUMOUCHEL, W.H. (*Massachusetts Institute of Technology*)
BERNARDO, J.M. (*Universidad de Valencia*)
FRENCH, S. (*University of Manchester*)
KADANE, J.B. (*Carnegie-Mellon University*)
LINDLEY, D.V. (*University College London*)
O'HAGAN, A. (*University of Warwick*)
SMITH, A.F.M. (*University of Nottingham*)
STROUD, T.W.F. (*Queen's University, Canada*)

REPLY TO THE DISCUSSION

The roles of inductive modelling and coherence in Bayesian statistics

TOM LEONARD*

Queen's University, Kingston, Ontario

University of Warwick

SUMMARY

The role of the inductive modelling process (IMP) seems to be of practical importance in Bayesian statistics; it is recommended that the statistician should emphasise meaningful real-life considerations rather than more formal aspects such as the axioms of coherence. It is argued that whilst axiomatics provide some motivation for the Bayesian philosophy, the real strength of Bayesianism lies in its practical advantages and in its plausible representation of real-life processes. A number of standard procedures, e.g. validation of results, choosing between different models, predictive distributions, the linear model, sufficiency, tail area behaviour of sampling distributions, and hierarchical models are reconsidered in the light of the IMP philosophy, with a variety of conclusions. For example, whilst mathematical theory and Bayesian methodology are thought to prove invaluable techniques at many local points in a statician's IMP, a global theoretical solution might restrict the statistician's inductive thought processes. The linear statistical model is open to improvement in a number of medical and socio-economic situations; a simple Bayesian alternative related to logistic discrimination analysis often leads to better conclusions for the inductive modeller.

Keywords: COHERENCE; INDUCTIVE MODELLING; AXIOMS; BAYES FACTOR REGRESSION; DISCRIMINATION; SKEWED-NORMAL; MULTI-PARAMETER ESTIMATION

1. THE RÔLE OF BAYESIANISM IN THE REAL WORLD

An overwhelming majority of practical statistical problems fall into a particularly general category. The statistician S is frequently required to investigate a real-life process R_ℓ and to extract some meaningful conclusions from his investigation. He might for example be faced with a large-scale set of medical data, and a team of medical experts, and might wish to assist in the diagnosis of the main causes of a particular disease. Alternatively, he may be

* Now at University of Wisconsin-Madison

concerned with a production process, e.g. for synthetic fibres, and be required to either forecast future output or to help detect ways in which the process can be improved. As a third example, he may be working in an educational testing environment, with the task of identifying students who could usefully attend particular colleges.

The Bayesian philosophy provides an excellent conceptual background for S's investigation of R_ℓ. As each fresh piece of information about R_ℓ becomes available to S, he is able to use it to refine his overall appreciation of R_ℓ. Whilst he might try to do this in a completely intuitive way, Bayesianism will frequently assist him in crystallising his complex thought processes, and in keeping his ideas on a sensible track.

It is one of the main themes of this paper that, whilst mathematical theory and Bayesian methodology play valuable *local* rôles in helping to clarify S's thought processes at a variety of points in his investigation of R_ℓ, they should not be expected to lead to a meaningful *global* solution to the problem of how S should approach his overall investigation of R_ℓ.

Even if it were technically possible to construct a feasible 'global' theory, we feel that such a solution would be inevitably restricted by the boundaries of its own assumptions, and could serve to constrict the inductive reasoning which is so vital to our understanding of the real world, and which no deductive theory can properly represent. For example, it is frequently the appearance of something completely unexpected which leads to new discoveries and important innovations. If our theory were insufficiently innovative to incorporate the possibility of all unexpected occurences in advance, then it might merely serve to disguise the potential discovery in a manner contrary to the general principles of science.

Similarly, if S wishes to develop a mathematical model as a device for extracting real-life conclusions from the data, then theory on its own would need to assume an enormously superhuman capacity to always select an inductively sensible model from a set of alternatives specified in advance. By examining the data, getting a good feel for its properties and its background, and interacting between the data, the client, tentative models and analyses, and possible real-life conclusions, S will often be able to use his inductive thought processes to help him to extract rich and meaningful conclusions from the data, which might well have remained undiscovered if he had followed a more formal philosophy.

In this *inductive modelling* process (IMP) which should be viewed as the basis of statistical practice. Whilst mathematical theory and Bayesian methodology will provide invaluable assistance at many local points of IMP, a more global concentration on these aspects may well lead S to either work in a theoretical vacuum or to become restricted by theoretical formalisms.

2. FORMAL AND INFORMAL JUSTIFICATIONS OF BAYESIANISM

The statistician S will typically need to convince his client of the possible benefits of Bayesian procedures when compared with other e.g. frequentist procedures. How should he seek to do this? It seems to us that S should simply try to convince his client that (a) Bayesianism often leads to a much more reasonable conceptual representation of aspects of R_ℓ and that (b) when applied to local problems, Bayesian methodology frequently leads to superior practical results (e.g. (i) multi-parameter estimation, (ii) problems involving nuisance parameters).

A number of authors (e.g. De Groot, 1970, pp. 71-76; Savage, 1954 and de Finetti, 1975) have devised axiom systems which, if acceptable to S, lead to the conclusion that he must act like a Bayesian, e.g. by representing his information by a probability distribution. Whilst some Bayesians might view such axiom system simply as a helpful description of the Bayesian approach, others (e.g. the Lindley-Smith-Dickey-Hill school) view such 'axioms of coherence' as compelling reasons for acting like a Bayesian and might even be tempted to employ such extremely appealing verbal arguments as 'Well, if you don't act like a Bayesian then you must be incoherent!'.

Most such axiom system seem acceptable from a formal point of view and it would appear sensible to act like a Bayesian if R_ℓ were simple enough to permit this. However, whilst many arguments in favour of Bayesianism based upon axiomatics possess substantive appeal, and whilst it would be pleasant if the axiomatic justifications turned out to possess a firm scientific basis, they may provide as convincing a justification as we might have hoped for.

In discussing ways of justifying Bayesianism, it might be useful to consider a particular set of axioms in detail. The set described by DeGroot is probably one of the easiest to follow; it is not confused by any notions of betting and its assumptions are similar in strengh to those suggested by most previous authors. They appear to have been suggested by DeGroot himself more as a description of the Bayesian approach then as a justification of it; they are related to the work of Villegas (1964).

The axioms consider a space Ω (which could for example be viewed as the space of all possible states of R_ℓ) with a sigma-field $_a$ of events, where any two elements A and B of $_a$ can be compared using the notation $A < B$ to indicate that S considers B to be more likely than A, $A \sim B$ to indicate his opinion that A and B are equally and $A \leq B$ to indicate that either $A > B$ or $A \sim B$; For the final axiom we require the definition

Df.: A quantity X is a *uniformly distributed random variable on the interval* [0,1] if for any two sub-intervals I_1 *and* I_2 of [0,1], $[X \in I_1] \leq [X \in I_2]$ if, only if, $\lambda(I_1) \leq \lambda(I_2)$, where $\lambda(I)$ denotes the length of the interval I.

The five 'axioms of coherence' are

Axiom 1: For any A and B, either $A<B$, or $A>B$, or $A \sim B$.

Axiom 2: For any A_1, A_2, B_1, and B_2, such than $A_1 \cap A_2 = B_1 \cap B_2 = \phi$ and $A_i \leq B_i$ for $i=1,2$, then $A_1 \cup A_2 \leq B_1 \cup B_2$. If in addition either $A_1 < B_1$ or $A_2 < B_2$ then $A_1 \cup A_2 < B_1 \cup B_2$.

Axiom 3: For any A, $\phi \leq A$. Furthermore $\phi < \Omega$.

Axiom 4: If $A_1 > A_2 >$ is a decreasing sequence of events and B is some fixed such that $A_i \geq B$ for $i = 1,2,\ldots$. then $\bigcap_{i=1}^{\infty} A_i \geq B$.

Axiom 5: There exists a uniformly distributed random variable X on interval [0,1].

The first three of the above axioms would probably seem reasonable to statisticians of most philosophies. Attempts should therefore be made to satisfy them, at least approximately, in local situations where an overemphasis would not detract S from the main purpose of his IMP, e.g. to induce real-life conclusions from the data. They lead to an approach described by DeGroot as 'relative likelihood', but do not in themselves give the slightest hint of a probability distribution on Ω.

The fourth axiom may be viewed as a regularity condition which ensures that the probability distribution, induced by Axiom 5, is countably additive rather than finitely additive.

The fifth axiom and its implications are of paramount importance. It introduces the notion of an auxiliary experiment (e.g. the spin of a roulette wheel) which yields an (objectively) random number X in the interval [0,1]. The statistician S is expected to be able to compare events in Ω with events on [0,1]. DeGroots's theory then leads to the construction of a unique probability distribution over Ω which represents S's feelings about elements of Ω and hence provides us with the result that S is actually acting like a Bayesian.

Implicit in DeGroot's formulation is the assumption that the first four axioms relate to any (measurable) subsets of the union of Ω and [0,1] as well as of Ω itself. It seems obvious that it is this implicit axiom (*5a*) which is primarily responsible for inducing the probability distribution on Ω since it maps subsets of Ω into the interval [0,1] in a mathematically rigorous way. It also seems that axiom *5a* is virtually as strong as the final result and that we are therefore very nearly saying "if you want to act like a Bayesian then you must act like a Bayesian"!.

Consequently, whilst axiom *5a* and the final result both possess considerable inductive appeal for Bayesians, the axioms do not in themselves appear to add anything beyond a useful interpretation of Bayesian thinking, in terms of an auxiliary experiment. The axioms should certainly never be used as a justification for Bayesianism or as a device for convincing non-experts. It would be more reasonable to refer to the justifications discussed in (*a*) and (*b*) above.

When S is engaged in his IMP, he may find it useful to employ the ideas of coherence as a conceptual background, to help him think upon Bayesian lines. If however he sticks too closely to axiomatics then he may lose sight of the primary objective of his investigation e.g. to extract real-life conclusions from the data. He should not permit coherence to restrict his creative and innovative ideas and he should concentrate more closely on appreciating the practical situation at hand. A good inductive appreciation of R_ℓ with a background culture of Bayesian coherence is to be preferred to an over rigid approximation to coherence and a lack of appreciation of R_ℓ.

The philosophy of coherence may be viewed in similar spirit to the ideas of Birnbaum (1962), which probably comprised one of the best single contributions to theoretical statistics. Birnbaum proved that the sufficiency principle and the conditionality principle together imply the likelihood principle, a far-reaching result which enables the purist to disregard many frequentist procedures integrating across the sample space.

The conditionality principle possess similar appeal to Axiom *5a* described above, and whilst acceptable in an idealistic sense, it is primarily responsible for Birnbaum's result that statisticians should follow the likelihood principle. When S is engaged in his IMP he may find it too restrictive to stick rigidly to the conditionality principle. For example, a responsible S would, as a general norm, obtain a good feel for his data before inducing a family of sampling distributions for his observations.

A related practical difficulty associated with Birnbaum's approach is that it is a conditional philosophy, given the truth of an underlying model for the observations. Any debate which conditions on the truth of an underlying model may be well wide of the target in the light of the philosophy "All sampling models are ultimately wrong and should simply be introduced as subjective, mathematical devices, in order to induce real-life conclusions from the data". This philosophy is an essential ingredient of our whole concept of *IMP*; it seems to provide us with one of the few sensible ways of engaging in a modelling process, and immediately detracts attention from philosophies which depend upon the truth of an underlying model.

3. JUSTIFIFYNG REAL-LIFE CONCLUSIONS

Once S had induced a real-life conclusion from the data and his appreciation of R_ℓ, he might wish to compile evidence in support of his conclusion, so that he can convince his client and other experts that it is both viable and meaningful. For example, in a paper to be published elsewhere, (but discussed in the verbal presentation of this material), Leonard, Low and Broekhoven (1978) describe a conclusion which is not in immediate concurrence with existing medical opinion. They have found that, whilst a high risk of fetal asphyxia in babies does not in fact appear to be noticeably associated with prematurity it does appear to be strongly associated with babies who possess a much lower birthweight than might be expected, for a given degree of prematurity.

These are several possibilities open to S, for example.

(a) To test his underlying model against the data, using a conventional significance test.

(b) To informally evaluate his model and conclusions by checking them out against future observations.

(c) To informally check out his real-life conclusions against the present data set, look for patterns in the conclusions, and consider their status in connection with existing scientific knowledge on related topics.

(d) To discuss his conclusions in detail with his client, to see if they fit in sensibly with his existing views, or whether the latter can be sensibly modified to accomodate his conclusions.

(e) To refer to the level of expertise of his own inductive judgement.

I feel that (*a*) should not be regarded as completely adequate, though significance tests may be useful as intuitive devices. Firstly, situations could be envisaged where the model is inadequate, but the specific conclusions are still viable. For example, a very tentative model could be used to stimulate plausible creative ideas by S, or the real-life conclusions might only depend upon particular aspects of the model. More importantly, significance tests do not appear to possess too much formal justification. For example, Leonard (1979) shows that for large sample sizes, significance levels may be sensibly replaced by value depending on the sample size. For further discussions of significance testing see Leonard and Ord (1976), and Leonard (1977 and 1978).

The alternative (*b*) appears to provide a useful check. However, the number of future observations will typically be finite and probably never particularly large. Also, by the time they have been collected R_ℓ will probably have evolved into an updated situation, and the usefulness of any underlying model undetermined. Just as the practical viability of the theoretical concept of consistency may be critically exposed in the context of the philosophy "the greater the amount of information the greater the chance of contradiction (of the original model)", the usefulness of predictive validation seems affected by the possible deviation of future observations from the situation currently at hand, whenever there are enough future observations to provide a case for a through validation.

Whilst (*c*) and (*d*) also provide useful checks, we feel that in the last analysis S can only refer to (*e*) and recognise that both R_ℓ and his investigation of it are basically subjective. He can only really attempt to justify his conclusions by simply indicating that he has carried out a subjective and honest investigation of R_ℓ and that his conclusions appear to be sensible.

We have thus arrive at the straightforward proposition that statistical practice is a subjective process which is highly dependent upon the expertise, honesty, and experience of the statistician, just as the practice of, say, medicine, law, psychology, economics, and indeed most branches of science, is also subjective and highly dependent upon similar qualities of experts in those areas.

In particular, the statistician will only be able to adequately complete his *IMP* if he possesses the mathematical skills and level of creativity which will carry him through the numerous local and innovative procedures which *IMP*'s typically require. People working from a "cookbook" of recipes will typically find difficulty with *IMP*'s and should therefore be discouraged from playing a leading rôle in large-scale investigations. The ultimate success of Bayesian statistics will depend upon whether we can bridge the gap between theory and practice and link theoretical innovation with practical relevance.

4. CHOOSING BETWEEN DIFFERENT SAMPLING MODELS

During his *IMP*, S may wish to use a formal Bayesian procedure to help him to measure his opinions about a finite number of sampling models. A number of authors (e.g. Dickey 1975, and Harrison and Stevens, 1976) have proposed a general approach to this problem, based upon sharp hypotheses and mixed models. However, whilst Schwarz (1978) has developed an approximate method for large sample sizes, which does not depend upon the choice of prior distribution, the general approach experiences some technical difficulties for smaller sample sizes. When more than two or three models are involved in the mixture it also appears to us to place too much emphasis on the search for a 'true' sampling model, and to be somewhat overcomplex and insufficiently motivated towards the extraction of meaningful real-life conclusions from the data. An informal consideration of alternative models in the light of real-life aspects may be more appropriate, i.e. we view the Bayesian mixed model approach as often assuming too much of a 'global' nature to provide an inductively useful service for S.

Suppose that S wishes to choose between a binomial sampling model with probability θ and sample size n for a frequency x and an alternative sampling model with probability mass function $p_0(x)$. For simplicity, we suppose that $p_0(x)$ is completely specified; assume also that whenever the binomial sampling model holds, θ possesses the beta prior distributions.

$$\pi(\theta\,|\,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}(1\text{-}\theta)^{\beta-1}(0\leq\theta\leq 1;\, 0<\alpha,\beta<\infty) \qquad (1)$$

Following the general approach referenced above, the posterior probability that model p_0 holds, given that either p_0 or the binomial sampling model holds, is then denoted by

$$\phi_0 = \frac{\phi R_x}{\phi R_x + 1} \qquad (2)$$

where ϕ is the corresponding prior probability, and R_x is the 'Bayes factor' which satisfies

$$R_x = p_0(x)/D(\alpha,\beta) \qquad (3)$$

where

$$D(\alpha,\beta) = \frac{\Gamma(n+1)\,\Gamma(\alpha+\beta)\,\Gamma(\alpha+x)\,\Gamma(\beta+n\text{-}x)}{\Gamma(x+1)\,\Gamma(n\text{-}x+1)\,\Gamma(\alpha+\beta+n)\Gamma(\alpha)\Gamma(\beta)} \qquad (4)$$

Whilst (2) provides a formal and coherent Bayesian solution to this problem, it is so sensitive to the choice of prior distribution for θ that it would be viewed as impractical in many situations. Suppose, for example, that α is moderately large and is increased by a single hypothetical prior observation to $\alpha + 1$. Note from (4) that

$$\frac{D(a+1,\beta)}{D(a,\beta)} = (\varrho p + (1\text{-}\varrho)\,\xi)/\xi \qquad (5)$$

where $p = x/n$, $\xi = \alpha/(\alpha + \beta)$, and $\varrho = n/(\alpha+\beta+n)$.

Therefore under our minor adjustment to the prior the Bayes factor in (3) should be divides by the quantity in (5), which will always lie between p/ξ and unity. For example, with the proportion p equal to 9/10 and the prior mean ξ equal to 1/10, the divisor could be as high as 9, radically, affecting the posterior probability in (2).

Paradoxically the sensitivity is at its greatest at $n \rightarrow \infty$, with p, α, and β fixed, so that $p \rightarrow 1$. In this case, the Bayes factor in (3) will tend to either zero or infinity irrespective of the prior, but the rate of convergence will become particularly sensitive, as increasing α by unity is equivalent to dividing the Bayes factor by the maximum possible value of p/ξ.

The sensitivity described above is not unique to the present special case. For example, Lindley (personal communication) has informed us that there are a further sensitivity problems when investigating whether or not to take observations to be normally distributed. Other problems concerning this type of approach are discussed by Atkinson (1978).

We are drawn to the viewpoint that it may be inductively more sensible to choose a sampling model by considering various aspects of R_ξ, and the data, and by generally following the philosophy outlined in the last paragraph of section 1 rather than by referring to a coherent Bayesian procedure with possible misleading conclusions. Note that sensitivity problems occur very generally in a number of other areas of Bayesian estimation and inference; some of these will be discussed in forthcoming publications by J.Q. Smith and J. Kadane.

5. THE RÔLE OF BAYESIAN PREDICTIVE DISTRIBUTIONS

A number of authors, e.g. Aitchison and Dunsmore (1975) view predictive distribution as playing a leading role in Bayesian methodology. It is our own view that whilst many standard predictive distributions, e.g. based upon conjugate prior distributions, play a role in idealised situations where the sampling model and prior distribution can be precisely specified, they may be of more limited importance when S is engaged in the practical details of his *IMP*. This conclusion is primarily based on the following reasons:

(a) Many predictive distributions can be as sensitive to the choice of prior as the Bayes factors discussed in section 4. For example, if (1) provides the posterior distribution for a probability θ, then the quantity $D(\alpha,\beta)$ in (4) is just the predictive probability that a binomial frequency, with probability θ and sample size n, is equal to x. Therefore if α is increased to $\alpha + 1$, this predictive probability will multiplied by a factor of up to p/ξ where p and ξ now respectively denote the predicted proportion x/n and the posterior mean $\alpha/(\alpha+\beta)$.

(b) The statistician S will typically remain uncertain about the correctness of his sampling model, and many conventional predictive distributions fail to take account in this uncertainty.

Suppose, for example, that we analyse a set of data which appear to be roughly normally distributed, that the practical situation (e.g. quality control) requires us to predict the probability that a further observation will be negative, and that the proportion of negative observations is 0.27. We then derive a standard predictive t-distribution under normal and conjugate assumptions and find that our predictive probability, conditional on our choices of sampling and prior models is 0.15. The latter is however a highly conditional probability and it might therefore be highly misleading to quote it as a useful result. Whilst our intention might suggest that a better (subjective) predictive probability lies between 0.15 and 0.27, many formal procedures for judging it more precisely would also be highly dependent upon any assumptions made.

Our general philosophy that "all sampling models are ultimately wrong" (see the last paragraph of section 3) leads us naturally to the philosophy that "all predictive distributions based upon particular sampling models are ultimately wrong". Conclusions based upon them sould be treated with caution.

We view many conventional predictive distributions as a bit on the over-formalistic side; indeed many standard predictive distributions do not obviously lead to any further inductive understanding of R_ℓ beyond that already provided by the sampling distributions from which they are generated. Many probabilities calculated from predictive distributions can only be considered to lead to reasonable practical predictive probabilities if these fit in closely with raw probabilities calculated from the data, or if there is some further inductive reason for using them. However, an alternative type of predictive distribution yielding greater scope to the inductive modeller will be discussed in section 7.

6. SOME PRACTICAL ADVICE ON THE LINEAR MODEL

We now discuss some practical aspects of the linear model, and consider dependent variables y_i satisfying

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \qquad (i=1,\ldots,m) \tag{6}$$

but where, for $q \leq p < m$, $x_{i1},\ldots,x_{iq}$ are statistical observations rather than fixed constants, and where $x_{iq+1},\ldots,x_{ip}$ are functions of $x_{i1},\ldots,x_{iq}$. The y_i could denote the salaries of m individuals, and $x_{i1},\ldots,x_{iq}$ could measure socio-economic factors relating to these individuals. Alternatively, y_i could represent blood pressure, with $x_{i1},\ldots,x_{ip}$ measuring q different medical symptoms.

It is my practical experience, and the general experience of colleagues in a consulting capacity, that there are a large number of practical situations where the underlying assumptions of the linear model seem appropriate, but where a modelling procedure of this nature turns out to be rather inadequate. This is particularly true of many socio-economic and medical data sets whenever there is a large amount of random fluctuation between the vectors $\mathbf{x}_i = (x_{i1},\ldots,x_{iq})$. In such circumstances it is often virtually impossible to arrive at any sensible model of the form defined in (6), whatever functional forms are chosen for $x_{iq+1},\ldots,x_{ip}$, and whatever estimation procedures (e.g. least squares, weighted least squares, or Bayes) is employed.

The data sets referred to might be viewed as possessing insufficient information to present the possibility of useful conclusions. Alternatively, a novice might feel tempted to add more and more explanatory variables in attemption to obtain a meaningful model. However, the simple Bayesian procedure described in section 7 and relating to logistic discrimination analysis very frequently leads to useful conclusions which would often be missed by the linear statistical modeller.

For a number of data set of this type, we have experienced a residual sum of squares which remains steadfastly close to the total sum of squared for virtually any model specification of the type defined in (6). This is because the $\mathbf{x}_i$ vectors are subject to so much random variation that it is almost impossible to use any set of fitted values to provide reasonable numerical predictions of the dependent variables i.e. the information content of the data is not of a predictive nature. Whilst inductive conclusions might still be available via the linear model, they will frequently be of limited scope owing to the extreme inadequacy of the model. For example, difficulties (b) involving predictive distributions, as discussed in section 5, will be highlighted in this context.

Whilst linear models present difficulties when the information content of the data is not a *predictive* nature, the same data sets often contain some very worthwhile information of a *probabilistic* nature which can be extracted via the methodology of section 7. The latter will also be enable S to model terms corresponding to $x_{iq+1},\ldots,x_{ip}$ in a direct (rather than, say, stepwise) fashion; for example it will help him to induce the presence of any complicated interaction effects without needing to engage in a long search.

Consider for simplicity the special case where $q = 1$ and $x_{i1} = x_i$. Suppose that the points (y_i, x_i) are plotted on a scatterdiagram for $i=1,\ldots,m$. Whilst these points will seldom lie close to any particular curve for the type of data set under consideration, the frequencies of y's falling in any particular intervals will often change in a meaningful way as x increases, as long as this interval is chosen to be wide enough. Therefore, whilst fitted values under any linear model might give poor numerical predictions of the y_i, it might be possible to use the data to help predict probabilities for intervals in which, say, a further observation y_{m+1} might lie, so that the data possess a quality of a probabilistic rather than a predictive nature. In other words, knowledge of a further explanatory variable x_{m+1} might affect S's probabilities about y_{m+1} but not provide him with enough information to be able to numerically predict y_{m+1} to any degree of accuracy.

It is my experience that data sets possessing information of this probabilistic rather than predictive nature a occur frequently in socio-economic and medical contexts, and that the linear model frequently possesses very limited scope for the analysis of such data sets. For example, many applications of the linear model to economics, sociology and medicine, might benefit from further consideration.

7. A BAYESIAN *IMP*

In the situation discussed in the previous section, where the information content is of a probabilistic rather than a predictive nature considerable headway may often be made upon categorising the dependent variable y. This will clearly lead to some loss of sampling information, but the loss need not be at all substantial, (owing to the highly random nature of the explanatory variables), as long as the dependent variable is categorised in a sensible way. For example, in the medical context of Leonard, Low and Broekhoven, three categories, referred to as 'low'; 'medium' and 'high', with the boundary points based upon further medical considerations, were adequate to permit the extraction of some meaningful conclusions from the data.

If the dependent variable is split into s categories, then the vectors $x_1,\ldots,x_m$ are effectively sectioned into s subpopulations $\Lambda_1,\ldots,\Lambda_s$, where the elements of Λ_j are those $\mathbf{x}$'s for which the corresponding y lies in category j. We let $n_1,\ldots,n_s$ denote the numbers of $\mathbf{x}$'s falling in the respective subpopulations $\Lambda_1,\ldots,\Lambda_s$.

Since the $\mathbf{x}$'s are themselves vectors of statistical observations, the $\mathbf{x}$'s in each sub-population Λ_j may be viewed as comprising a random sample from a distribution, say with density f_j ($\mathbf{x}$). The form of this density may be inductively modelled by S in the light of the corresponding $\mathbf{x}$'s and his appreciation of $R\ell$. This provides a vital part of S's *IMP* in this context; he needs to model the s densities $f_1,\ldots,f_s$. Suppose now that S wishes to be able to predict probabilities for a further dependent variable y_{m+1}, given a further vector of explanatory variables $\mathbf{x}_{m+1}$. Then the probability that y_{m+1} falls into the j^{th} category, given that $\mathbf{x}_{m+1} = \mathbf{x}$ is given by

$$\text{prob}\,(\Lambda_j|\mathbf{x}) = \frac{\pi_j f_j(\mathbf{x})}{\Sigma_{k=1}^{s}\,\pi_k f_k(x)} \qquad (j=1,\ldots,s) \tag{7}$$

where π_j denotes the corresponding prior probability, However, in the absence of knowledge of $\mathbf{x}$, S will frequently be prepared to set. $\pi_j = n_j/m$ for $j=1,\ldots,s$, in which case we have

$$\text{prob}\,(\Lambda_j|\mathbf{x}) = \frac{n_j f_j(\mathbf{x})}{\Sigma_{k=1}^{s}\,n_k f_k\,(x)} \qquad (j=1,\ldots,s) \tag{8}$$

The formula in (8) may be applied in a simple way to data sets whose information content is of a probabilistic nature; it seems to fit in neatly with the concept of *IMP*. It provides a standard procedure for many regression problems which could be used as an alternative to analyses based upon the linear model.

Note that the expression on the right hand side of (8) plays the rôle of a regression function. We for example have

$$\begin{aligned}\log\,[\text{prob}\,(\Lambda_j|x)/\;\text{prob}\,(\Lambda_k|\mathbf{x})] &= \log\,(n_j/n_k)\\ &+ \log f_j(\mathbf{x})/f_k/(\mathbf{x})\end{aligned} \tag{9}$$

This result is employed in logistic discriminent analysis. For example, Anderson (1974) mentions that multivariate normal assumptions for the f_j lead to a quadratic discriminant of regression function on the right hand side of (9).

Under our general *IMP*, S is expected to simply induce $f_1,\ldots,f_s$ from the $\mathbf{x}$'s and $R\ell$. Our point is that no further modelling will then be required because appropriate substitutions in (8) will complete the specifications of the predictive probabilities. During this process, S will need to interact between scatterdiagrams of the $\mathbf{x}$'s in the different sub-populations and his other experience and he will therefore be able to take full account of the probabilistic-type information content of the data. This inductive modelling will enable him to obtain predictive probabilities via (8), By considering graphical plots of the latter against different explanatory variables he is then in a position to extract real-life conclusions from the data.

Note that the above *IMP* automatically models the form of the regression function and hence the presence of any interaction effects, even if these are of a complex nature. As a simple example, multivariate normal assumptions for the f_j lead to cross-product terms on the right hand side of (9), which may be viewed as the interaction terms in a logistic regression. They now become completely determined upon identification of the f_j, providing a much more straightforward modelling procedure, then, say, standard stepwise procedures for the linear model. For non-normal f_j the interactions can assume a much more complex nature, but S has a very straightforward way of inducing them.

We recommend replacing any unknown parameters in the f_j by suitable points estimates (e.g. maximum likelihood or Bayesian). This should be frequently superior to the coherent Bayesian procedure of integrating each f_i in (8) with respect to the corresponding prior distributions of the parameters, since the latter will suffer similar sensitivity problems to those discussed in sections 4 and 5.

There are a number of ways of checking the probabilities in (8) against the data set. For example, boundaries on **x** could be determined for each j such that prob $(\Lambda_j|\mathbf{x})$ is greater than a specified value. Then the proportions of actual **x**'s falling inside these boundaries could be enumerated, and they will all ideally be greater than the specified lower bound for the predictive probability. Added credibility will also be given to the *IMP* if the curves of prob $(\Lambda_j|x)$ against **x** evolve in a sensible way for increasing j.

The above approach has been found to yield practical conclusions in a variety of different situations, than would appear possible under a standard linear model approach. Similar methodology was employed by Leonard, Low and Broekhoven in their medical context.

8. THE SKEWED-NORMAL DISTRIBUTION

The statistical modeller is frequently faced with data with both a positive and negative tail, and which indicate a definite skewness. There are surprisingly few probability distributions in the literature for adequately modelling skew data when the latter are scattered on the whole real line. The following properties would however seem to be desirable for a family of two-tailed distributions which provide skew alternatives to say, the normal or t-distribution:

(i) A meaningful set of at least three parameters, with convenient functions of the parameters representing location, spread and skewness.

(ii) A useful symmetric distribution as a special case.

(iii) The property that whilst the two tails can be different they should be 'similar in nature', in the sense that different functional forms assumed for the tails might suggest a difference which was not exhibited by the data.

(iv) The form of the likelihood function, given n observations, should not permit the observations in one tail to unduly influence the estimated thickness of the other tail.

(v) Straightforward *ad hoc* and Bayesian estimation procedures for the parameters.

(vi) Easily tabulated interval probailities.

(vii) Reasonable regularity conditions for the density e.g. a continuous first derivative at all points.

All the above properties are satisfied by the *skewed-normal distribution*, with parameters μ, σ_1^2, σ_2^2, and density

$$p\,(x\,|\,\mu,\sigma_1^2,\sigma_2^2) = \begin{cases} \sqrt{(2/\pi)}(\sigma_1+\sigma_2)^{-1}\exp[-½\,\sigma_1^{-2}(x-\mu)^2] \text{ for } x \leq \mu \\ \sqrt{(2/\pi)}\,(\sigma_1+\sigma_2)^{-1}\exp[-½\,\sigma_2^{-2}(x-\mu)^2] \text{ for } x \geq \mu \end{cases} \qquad (10)$$

This distribution possess mode μ and probabilities $\sigma_1/(\sigma_1 + \sigma_2)$ and $\sigma_2/(\sigma_1 + \sigma_2)$ either side of the mode. Its technical properties, including a Bayesian analysis, will be reported in more detail elsewhere.

9. SUFFICIENCY, OUTLIERS AND COHERENCE

In many statistical problems, the existence of a sufficient statistic of small dimensions implies in effect that the sampling distributions is a member of the exponential family. Therefore any discussion of the inductive reasonability of the concept of sufficiency must be closely related to a debate on the adequacy of the exponential family of distributions.

The general concept of sufficiency could be criticised on the grounds that a sufficient statistic typically reduces the number of pieces of information we can extract from the data, i.e. from the sample size to the dimension of the sufficient statistic. The data are therefore reduced to a form where they can, say, only describe one or two aspects of the sampling distribution, e.g. location and spread, but may tell us nothing about, or even disguise, other important aspects of the sampling distributions, e.g. possible bimodality or thicker tails than might be experienced with the exponential family.

Consequently, in situations where we might wish a formal analysis to tell us as much as possible about the sampling distribution, the concepts of sufficiency and the exponential family of distributions do not seem to be completely adequate. The formal Bayesian could, for example, be tempted to refer to the interesting approach of O'Hagan (1979) and employ outlier-prone and outlier-resistant sampling distributions in an attempt to cope with outliers.

On the other hand, sampling distributions yielding sufficient statistics typically possess meaningful characteristics and meaningful parameters. They seem to fit in well with the concept of *IMP* since S should always examine the data carefully and get a good feel for its properties before inducing a sampling distribution. He could for example investigate bimodality and outliers intuitively rather than referring to the formalisms of a more complicated sampling model.

The statistician would probably do best to compromise between these two extremes. He could start off by referring to meaningful sampling distributions, with simple sufficient statistics, and to practical judgements of the data, with the objective of concentrating on the extraction of real-life conclusions from the data. However, he will sometimes find that his induction is unable to provide him with a clear enough picture. In this case slightly more complicated sampling distributions and an analysis taking formal account of further aspects of the data would sometimes be very useful.

As an example of the above approach, the skewed normal distribution in (10) is frequently applicable to (clearly unimodal) data with two tails. It can be employed as a useful device for locating the mode of the underlying distribution and for investigating its skewness. Its parameters are meaningful in this context; it provides a simple modification of a member of the exponential family. For example, when μ is known, statistics of the form

$$\sum_{i:x_i<\mu} (\mathbf{x}_i-\mu)^2 \text{ and } \sum_{i:x_i>\mu} (x_i-\mu)^2$$

are jointly sufficient for σ_1^2 and σ_2^2.

The skewed-normal distribution would clearly be inferior in a formal sense to a distribution with 't-type' tails if there we enough outliers in the data to suggest that its tails might be too thin. However, an adherent of *IMP* could still start off with the skewed-normal distribution and interact betwen tentative analysis based upon it, and the data, to see if the outliers affected the important real-life conclusions which could be induced from the data. For example, *S* could firstly try an analysis withouth the outliers, and then compare it with a further analysis with outliers present. Only if he convinces himself inductively that the outliers actually make a real difference should he consider a more formal (local) analysis based upon a complicated distribution with thicker tails. He is in this way able to increase his chances of extracting conclusions which might otherwise become confused by over complications.

The procedure outlined above is not obviously formally coherent, but we seem to have described a good example of a situation where a strict demand for formal coherence would appear to be inductively inappropriate.

10. MULTI-PARAMETER PROBLEMS AND PRIOR STRUCTURES

Consider next a general formulation where *S*'s $n \times 1$ observation vector $\mathbf{x}$ is though to possess a sampling distribution $f(\mathbf{x}|\theta)$ depending upon a $q \times 1$

vector $\theta = (\theta_1,\ldots,\theta_q)^T$ of unknown parameters. In such multi-parameter situations, *S* might be concerned about Stein-type effects and lack of smoothness of the maximum likelihood estimates, and might therefore wish to employ shrinkage estimates for the θ_i. (See, for example, a method proposed by Leonard, 1973, for smoothing the probabilities in a histogram).

Following a general procedure discussed by Leonard (1972), *S* might seek a $q \times 1$ vector $\alpha = (\alpha_1,\ldots,\alpha_q)^T$ of transformed parameters such that he is prepared to take the prior distribution of α to be multivariate normal, say with mean vector μ and covariance matrix $\mathbf{C}$. When μ and $\mathbf{C}$ are known a Bayesian shrinkage estimate for α is given by the posterior mode vector α, wich satisfies the equation

$$\left.\frac{\delta \log f(\mathbf{x}|\alpha)}{\delta\alpha}\right|_{\alpha=\tilde{\alpha}} = \mathbf{C}^{-1}(\tilde{\alpha}-\mu) \qquad (11)$$

For example, when all the elements of μ are equal to a scalar μ, and $\mathbf{C}$ is a scalar multiple of the identity matrix, the elements of α will be a priori *exchangeable* and (11) will roughly speaking provide Stein-type shrinkages of their maximum likelihood estimates towards a common value μ.

However, *S* is typically faced with the problems of choosing suitable special forms for μ and $\mathbf{C}$ and evaluating any hyperparameters appearing in these special forms (these forms may be referred to as *prior structures*). The situation will often be far too complex for *S* untangle if he confines himself to strictly coherent Bayesian procedures. We recommend that he should instead assess his prior structures by interacting between his prior feelings, possible special forms for μ and $\mathbf{C}$, tentative estimates obtained from (11), any real-life conclusions he can induce from these estimates, his overall experience of R_ℓ, and cooperation with his client.

S will find it difficult to assign specific values to any hyperparameters appearing in his prior structures. A typical prior structure may be expressed in the form $\mu=\mu(\lambda_1)$ and $\mathbf{C}=\mathbf{C}(\lambda_2)$, once *S* has induced the dependence of the mean vector and covariance matrix on hyperparameters λ_1 and λ_2. For any such prior structure under consideration *S* should estimate λ_1 and λ_2 from the data and any prior information which might be available. We are however rather uncertain about the existence of convenient prior information for hyperparameters in complex models like this, except in special cases or when the prior information is itself data based. It is generally much more straightforward to avoid complicated and possible confusing distributions at the second stage of the prior model, and to simply estimate λ_1 and λ_2 from the data by maximising their 'marginal likelihood'.

$$\ell(\lambda_1, \lambda_2, |\mathbf{x}) = E[f(\mathbf{x}|\alpha)] \tag{12}$$

where the expectation on the right hand side is with respect to α, given μ and $\mathbf{C}$.

In summary, S may induce the functional forms of $\mu(\lambda_1)$ and $\mathbf{C}(\lambda_2)$ by following the general philosophy of *IMP*, and may then estimate λ_1 and λ_2 via a data-based procedure. Obviously, particular practical considerations might lead to refinements of this scheme.

Whilst it would be difficult to demonstrate formal coherence of the above procedure, it seems likely to often prove useful in a real-life sense when compared with more complex coherent procedures.

11. NON-PARAMETRIC DENSITY ESTIMATION

The approach described by Leonard (1978) to the non-parametric estimation of a density fits with the philosophy of *IMP* since it enables S to allow for real-life considerations as part of theoretical local analysis. For example, a hypothesised density can be introduced as a prior estimate, then the theoretical method can be used to provide a posterior estimate which can be considered inductively by S, to see where it differs from his null hypothesis, and to consider whether these differences are due to real-life aspects. He could also try out different hypothesised densities as part of his *IMP*, and generally interact between his prior specification, his posterior results, and possibly meaningful conclusions. The approach seems to be more useful than many previous frequentist procedures based on kernel functions, since these tend to place a bit more emphasis on data-fitting, rather than on the diagnosis of meaningful conclusions.

Note that Leonard uses a prior and posterior likelihood approach rather than a strictly Bayesian approach since this avoids certain technical problems over function spaces. We in general see nothing wrong in following an alternative philosophy if it is based upon similar prior information and leads to similar conclusions.

12. DISCUSSION

The concept of coherence has played an invaluable theoretical rôle over the years by highlighting the inadequacies of many frequentist procedures. However, the Bayesian philosophy is now firmly established and accepted as one of the few viable theoretical approaches to Statistics. It should therefore now look beyond debates with other philosophies, and theoretical discussions on the foundations, and emphasise its practical viability in non-trivial contexts, e.g. large scale data sets where the client provides background information from his own discipline. When broader considerations are taken

into account the rôle of coherence no longer seems paramount, and much more emphasis should be place on the *IMP* aspects of statistics. Whilst existing coherent methodology is useful at a variety of local points of *IMP*, the theoretical structure should be kept to a level of intellectual complexity where it assists the statistician to induce real-life conclusions from the data.

ACKNOWLEDGEMENTS

The author wishes to thank Professor J.M. Dickey and L. Broekhoven for helpful advice and discussions.

REFERENCES

AITCHISON, J. and DUNSMORE, I.R. (1975). *Statistical Prediction Analysis*, Cambridge: University Press.

ANDERSON, J.A. (1975). Quadratic logistic discrimination. *Biometrika*, **65**, 39-48.

ATKINSON, A. (1978). Posterior probabilities for choosing a regression model, *Biometrika* **65**, 39-48.

BIRNBAUM, A. (1962). On the Foundations of Statistical Inference. *J. Amer. Statist. Assoc.*, 269-326.

DE FINETTI, B. (1975). *Theory of Probability* 1 New York: Wiley.

DEGROOT, M.H. (1970). *Optimal Statistical Decisions*. New York: McGraw Hill.

DICKEY, J.M. (1975). Bayesian alternatives to the *F*-test and least squares estimation in the normal linear model. In *Studies in Bayesian Econometrics and Statistics* (Fienberg and A. Zellner, Eds.), 515-554. Amsterdam: North Holland.

HARRISON, P.J. and STEVENS, C.F. Bayesian Forecasting (with Discussion). *J. Roy. Statist. Soc. B* **38**, 205-247.

LEONARD, T. (1972). Bayesian Methods for Binomial Data, *Biometrika*, **59**, 581-589.

— (1973). A Bayesian Method for histograms. *Biometrika*, **60**, 297-308.

— (1977). A Bayesian approach to some multinomial estimation and pretesting problems, *J. Amer. Statist. Assoc.* **72**, 867-874.

— (1978). Density estimation, stochastic processes and prior information (with Discussion) *J. Roy. Statist. Soc. B.* **40**, 113-146.

— (1979) Why do we need significance levels? *M.R.C. Tech. Report.* University of Wisconsin-Madison.

LEONARD, T., LOW, J.A., and BROEKHOVEN, L. (1978). Assessing the risk of fetal asphyxia. *STATLAB Tech. Report.* Kingston, Ontario: Queen's University.

LEONARD, T. and ORD, J.K. (1976). An investigation of the *F*-test as an estimation shortcut. *J. Roy. Statist. Soc. B* **38**, 95-98.

O'HAGAN, A. (1979). On outlier rejection phenomena in Bayes inference. *J. Roy. Statist. Soc. B* **41**. 358-367.

SAVAGE, L. (1954). *The foundations of Statistics*. New York: Wiley.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-4.

VILLEGAS, C. (1964). On qualitative probability Γ-algebras. *Ann. Math. Statist.* **35**, 1787-1796.

Local and regional coherence utility assessment procedures

M.R. NOVICK; D.F. DEKEYREL
and
D.T. CHUANG
The University of Iowa

SUMMARY

Novick and Lindley (1978, 1979) have dealt with the use of utility functions for applications in education and have advocated the use of the standard gamble (von Neumann and Morgenstern, 1953) elicitation procedure with the addition of coherence checking using overspecification and a least squares fit. In this procedure utilities are inferred from probability judgements offered by the assessor. This paper describes local and regional coherence procedures which seek utility coherence in successive restricted domains of the parameter space as preludes to overall coherence checking. These procedures and some others are viewed as possible ways of avoiding anchoring and certainty effect biases found in earlier fixed probability methods, and presumably present in current fixed state procedures.

Keywords: UTILITY ASSESSMENT, COHERENCE

INTRODUCTION

Earlier approaches to utility assessment (Mosteller and Nogee, 1951, Schlaifer, 1959, 1971; Raiffa and Schlaifer, 1961; Keeney and Raiffa, 1976; and so on) have been based on the use of fixed probability (FP) assessment procedures in which utilities are elicited directly, through successive bisections of the parameter space. It has been suggested (Mosteller and Nogee, 1951) that such procedures are easier to use because subjects are more familiar with the quantity for which the utility function is desired than they are with probabilities, which they are required to state in the standard fixed state (SFS) procedure.

Although it was originally thought by psychologists that utility theory would prove useful as a *descriptive* model (Swalm, 1966, etc.), much criticism has recently been levied against its use in that capacity. As principal critics, Kahneman and Tversky (1978) have proposed an alternative descriptive model. The main basis for their criticism is that the phenomenon described by Tversky (1977) as the *certainty effect* results in preferences that violate the substitution axiom or expected utility hypothesis of utility theory. This axiom (hypothesis) states that preference order is invariant over probability mixtures and is formally equivalent to the assumption that there is no positive or negative utility for the act of gambling itself. Specifically, the *certainty effect* is the phenomenon that the utility of an outcome seems greater when it is certain that when it is uncertain. This effect can be observed when subjects are presented with a choice between a for-sure and a chance option, the choice appearing in the standard gamble, regional coherence, and local coherence assessment procedures to be described in this paper.

Utility theory as considered here is used as a *normative* model rather than as a *descriptive* model; however, it is still important to consider the certainty effect because Tversky (1977) has shown that even when subjects were told that their preferences violated utility theory, they were not inclined to change them (see also Kahneman and Tversky, 1972). This brings into question the reliability (coherence) and bias-free character of utililty assessment procedures obtained through both fixed state and fixed probability methods and the value of those procedures in helping decision makers be more coherent. However, it should be pointed out that the gambles studied by Kahneman and Tversky and those studied by Novick and Lindley were somewhat different and that the latter authors also provided incoherence resolution procedures.

In another paper, Tversky and Kahneman (1974) described several heuristics used by persons in assessing probabilities and the biases to which they could lead. Of particular interest is the *anchoring and adjustment heuristic*, which Spetzler and Staël von Holstein (1975) have shown can reduce the reliability with which the bisection technique used by fixed probability models elicit utilities. This heuristic is the phenomenon whereby the most readily available piece of information often forms an initial basis for formulating responses from which subsequent responses are then adjusted. Since adjustments from this basis are often insufficient, a central bias results. According to Slovic (1972, the *anchoring and adjustment heuristic* is a natural strategy for easing the strain of integrating information. The anchor serves as a register in which one stores first impressions or the results of earlier calculations. Why adjustments from the anchor are usually insufficient, though, is unclear. Slovic advances two hypotheses to explain the insufficient adjustment. First, people may stop adjusting too soon because they tire of the mental effort involved in adjusting. Alternatively, the anchor may take on a special salience, thus causing people to feel that there is less risk in making estimates close to it than in making estimates that deviate far from it. According to Spetzler and Staël von Holstein (1975), experimentation has

shown that subjects tend to produce a central bias when, in the fixed probability bisection method, they are asked first for the median for an uncertain quantity and then for the quartiles.

Later, in reviewing the role of man-machine systems in decision analysis, Slovic, Fischhoff, and Lichtenstein (1977) suggested that human factors such as the ways in which variations in instructions or informational displays affect people's performance are important and should be studied in more detail. Questions of complexity and representativeness of material seem to have substantial effect on assessors responses (Fischhoof, Slovic and Lichtenstein, 1977; Vlek, 1973). The study of such factors might lead to an assessment procedure that minimizes the judgemental basis and heuristics described earlier. This position was strengthned by the discussion of Fischhoof, Slovic and Lichtenstein (1979). A consideration of these ideas promoted the development of a new format introduced later in this paper.

Extensive previous work in this area has raised more questions concerning bias and coherence than it has provided answers. An apparently pessimistic mood prevails, not inappropriately, given the importance of the questions that have been raised (Hogarth, 1975; Slovic, 1975; and Fischhoff, Slovic and Lichtenstein, 1979). Nevertheless, the very extensiveness of this research must itself imply a high assessment for the product of the probability of resolving these difficulties and the value of this outcome. The position taken here is that bias and incoherence can be reduced if (1) elicitations are carefully fashioned in a Computer-Assisted Data Analysis (CADA) environment (Novick, Hamer, Libby, Chen and Woodworth, 1980), (2) assessors are aided in resolving incoherence, and (3) if the assessments concern states and actions that are meaningful and important to the assessor at the time the assessment is made.

Consider a variable θ and the utility function $U(\theta)$ for which assessment is required. In most applications θ will be a real variable, such as grade point average (GPA), but this is not necessary. Although the contrary assumption is sometimes made, it seems sensible to us to demand that a utility function be bounded and increasing.

There are two standard approaches to assessing a utility function: fixed probability and fixed state. In the former, the subject is presented with a gamble on two values, or states, θ_1 and θ_2 with a fixed probability-p, say, for θ_1 and $1 - p$ for θ_2 - and is required to choose an intermediate state θ_3 such that he/she is indifferent with respect to the gamble and θ_3 for sure. In applications, typically $p = 1/2$ because this gamble is easiest for assessors (subjects) to understand.

In the fixed-state method, the states θ_1, θ_2, and θ_3 are fixed, θ_3 still being intermediate between θ_1 and θ_2. The subject is required to state a probability,

p, such that he/she is indifferent between θ_3 for sure and the following gamble: θ_1 with probability p and θ_2 with probability 1-p. If θ_1 and θ_2 have utilities of 1 and 0, respectively, the gamble has expected utility p, the indifference probability assigned to θ_3.

In the fixed-state method, let us suppose that a number of states, θ_0, $\theta_1, \ldots, \theta_{N+1}$, are selected. We shall further suppose that these states are ordered in the sense that θ_j is preferred to θ_i whenever $j > i$; in particular θ_{N+1} is the best and θ_0 the worst state. Then the utility function $U(\theta)$ will be strictly increasing.

Without loss of generality, the utility for θ_{N+1} can be assigned the value 1 and that for θ_0 can be assigned the value 0, thus placing bounds on the utility values to be assigned to the various states. We must then find N such values: $U(\theta_1)$, $U(\theta_2), \ldots, U(\Theta_N)$. We first consider adjacent gambles, that is, a situation in which the subject is asked to compare the sure outcome $\theta_n(1 \leq n \leq N)$ against a gamble with possible outcomes θ_{n-1} and θ_{n+1}, representing, because of the ordering of the states, situations respectively worse and better than θ_n. Specifically, after a brief review of the meaning of probability, the subject is asked to state the probability p_n for θ_{n+1}, and consequently 1 - p_n for θ_{n-1}, that makes him/her indifferent with respect to the gamble and θ_n for sure. Writing $U(\theta_n) = u_n$ (so that $u_0 = 0$, $u_{N+1} = 1$) and equating the expected utilities for the two situations gives us

$$u_n = p_n u_{n+1} + (1 - p_n) u_{n-1}.$$

If this done for all n, $1 \leq n \leq N$, we have N equations in N unknowns and aside from exceptional cases, the utilities are uniquely determined. The solution is

$$u_{n+1} = G_n / G_N$$

for $0 \leq n \leq N$, where

$$G_n = \Sigma_{i=0}^{n} F_i, F_n = \Pi_{i=0}^{n} f_i, \text{ and } f_i = (1 - p_i)/p_i.$$

Suppose a subject has responded to the N question previously considered and, from the answers given, his/her utilities u_1, $u_2, \ldots, u_N$ have been determined. Suppose also that he/she is asked to consider a gamble that will yield either θ_{n+2} or θ_{n-2}, against ν_n for sure. Then the probability q_n, associated with θ_{n+2}, satisfies

$$u_n = q_n u_{n+2} + (1 - q_n) u_{n-2}.$$

For the fixed state standard gambles procedure the suggestion offered by Novick and Lindley is that to exploit coherence fully, we must ask for more probability assessments than are needed to calculate the utilities and then compare them for coherence. The idea of requiring the experimenter to give more than the minimum number of judgments in fitting a personal probability distribution has been used by Pratt, Raiffa and Schlaifer (1965) and has been exploited systematically both for the assessment of probabilities and utilities in the development of the Computer Assisted Data Analysis Monitor (CADA) (Novick, 1973, 1975). In the context of utility assessments, the idea has been used by Becker, DeGroot, and Marschak (1963) with fixed probability assessments, and we shall discuss this presently.

Experience shows us that assessors are almost always incoherent but readily attempt to resolve their incoherences when these are brought to their attention (cf. MacCrimmon, 1965). It may, however, be true that one kind of gamble (e.g., adjacent gambles) may introduce one kind of systematic bias and another kind (e.g., extreme gambles) may introduce a second kind of bias. Therefore, rather than just asking the subject to revise some of his/her assessments, Novick and Lindley (1979) suggest assisting the subject by providing a least squares fit in the log-odds metric for the N undetermined utility values. A computer program has been written to carry out the interrogation of the assessor and to perform the least squares fit and is available on the CADA Monitor (Novick, et. al., 1980).

In any comparison of fixed state with fixed probability assessments, the role of coherence seems to us to play a dominant role. Although subjects often prefer fixed probability assessments, especially when the probability is 1/2, exploiting coherence in this context is harder than with the fixed state procedure. For example, suppose, as usual, that a subject is asked for the certainty equivalent of a gamble, at even odds, on the best (θ_{N+1}) and worst (θ_0) states. Let his/her stated value be θ_m, say, having $u(\theta_m) = 1/2$. The subject is then asked for the certainty equivalents for even-odds gambles on (θ_0,θ_m) and (θ_m, θ_{N+1}). If these values are denoted by θ_1 and θ_2, respectively, then the utilities of θ_1 and θ_2 are $u(\theta_1) = 1/4$ and $u(\theta_2) = 3/4$. Finally, he/she is asked to consider an even-odds gamble on θ_1 and θ_2. But it is rather transparent that for coherence the result must be θ_m, so that the four judgments can scarcely be considered independent. In this field (as in other measurement fields) obtaining independent repetitions of the same assessment is hardly ever possible, thus the emphasis ought to be on independent assessments of related quantities. This, we feel, is more nearly achieved with the fixed-state assessments. The above discussion is taken with some condensation from Novick and Lindley (1979).

The question that must now be addressed is whether the incoherence resolution of the least squares method described in SFS above avoids the certainty and anchoring effects or whether better methods can be found. The remainder of this paper will be devoted to describing a refinement in the least squares SFS procedure and in describing three new procedures that more directly address these biasing effects.

A word concerning ease of response may be in order. Mosteller was certainly correct in saying that FP is easier than FS, and, indeed, without interactive conversational computing facilities an FS assessment procedure may well be unbearably difficult. With conversational computing, however, an FS procedures is bearable and there is no reason to believe the easier method is more bias free. Indeed, the contrary could be true.

In the current version of the SFS procedure on CADA, subjects are given situations consisting of a for-sure and a chance option on grade point averages in the range 0-4 and are asked for the probabilities that make them indifferent with respect to the two options in each situation (i.e., their indifference probabilities). The indifference probabilities for the fixed state gambles are elicited using one of two formats for presenting the gambles. Format two request a direct magnitude estimation as illustrated earlier. Format one asks for preferences for gambles or sure things for p values .1, .9, .2, .8, etc., or .9, .1, .8, .2, etc., with zeroing in on the indifferent point.

FORMAT ONE

3.0 p chance	indifferent = 0
2.5 for sure	for sure = 1
2.0 1-p chance	chance = 2
	restart = 3

Which would you prefer if p were .XX ?______. (This question was repeated using the following p values .1, .9, .2, .8, ... until p was found to be between .5 and .6, say. Then the questioning procedure used p values of .52, .58, .54, ... until the subject's indifference p had been determined).

FORMAT TWO

for	gamble		p that makes
sure	with prob p	with prob 1-p	you indifferent
2.5	3.0	2.0	?______

TABLE 1 *Formats for the SFS utility assessment procedure*

Format two, the direct probability assessment format is the one used by Novick and Lindley, (1979). Format one, the *ends-in procedure,* has been advanced as a method for avoiding anchoring. Since indifference points are typically between .2 and .8 any initial anchor (.1 or .9) is erased before any careful judgment must be made. Also, the starting values alternate between .1 and .9 thus avoiding any constant ordering effect. It is our as-yet-unsubstantiated belief that this format is both easier and less subject to anchoring than format two. This format is now used with several assessment procedures.

In order to avoid the documented biases of the certainty effect in utility assessment, a new procedure has been considered: the paired binary gambles (PBG) procedure. This procedure is illustrated in Table 2 bellow. The paired gambles in the table can be abbreviated as (1.5 3.0, 2.0 2.5).

Paired Binary Gambles

SITUATION 1	SITUATION	2
3.0 p	2.5	p
1.5 1-p	2.0	1-p

TABLE 2 *PBG procedure*

The ends-in format is used to elicit the probability that will make the subject indifferent with respect to the two situations (gambles). A least squares fit of the indifference probabilities can then be made and subjects can proceed as in the SFS procedure.

Although the PBG procedure is considered here as a fixed state procedure, it has previously been used in a fixed probability paradigm (Kneppreth, Gustafson, Leifer, & Johnson, 1974). Suppes and Walsh (1959) have considered such gambles strictly in the sense of determining preferences between the two situations, without eliciting either indifference probabilities or equivalance points.

The obvious hope is that the PBG procedure will avoid the certainty effect because the comparison is between two sets of gambles, and thus does not involve the for-sure option. We have used PBG in some informal assessments but have not yet been convinced of its usefulness. First, it is difficult even for experienced subjects. Fatigue and boredom are definite problems. We are not sure that there is not a bias in that one situation always compares two adjacent states while the other always describes two states twice removed. We have not discarded this procedure, but we feel that refinements may be necessary if it is to be useful.

Next we define the Regional Coherence (RC procedure). In the RC procedure, indifference probabilities are elicited separately for two SFS gambles using the ends-in format. Subjects are then presented with a table showing the initial gambles (situations 1 and 2 with their indifference probabilities) and two additional gambles (situations 3 and 4). They are told that their initial responses imply certain specific indifference probabilities for the two new gambles. Table 3 illustrates the latter part of this procedure.

	Situations			
	1	2	3	4
p-chance	1.50	2.00	2.00	2.00
for sure	1.00	1.50	1.00	1.50
1-*p* chance	0.50	1.00	0.50	0.50
	$p=.53$	$p=.58$	$p=.40$	$p=.75$

TABLE 3 *RC Procedure*

Assessors are then given the opportunity to change the indifference probabilities, two at a time, until they are indifferent in all four situations. They choose the two situations for which they wish to change the indifference probabilities and the magnitude estimation format is then implemented to generate the revised probabilities.

The final procedure is called the local coherence (LC) procedure. This procedure presents subjects with two types of hypothetical choice situations: (1) a for-sure and a chance option (the standard gamble) and (2) two chance options. The ends-in format is used to elicit an indifference probability for the first situation, after which the subject is told that that response implies that he/she should be indifferent with respect to the two options in situation 2. Note that the subject only specifies the indifference probability for the standard gamble. Table 4 below illustrates this procedure. The probabilities for the second situation are uniquely determined by that specification.

SITUATION 1		SITUATION 2		
		option one		option two
4.00	.75 chance	.19	4.00	---
3.00	for sure	---	3.00	.25
1.00	.25 chance	.81	1.00	.75

TABLE 4 *LC Procedure*

If the subject is not indifferent in both situations, he/she modifies the situation 1 indifference probability and then is again presented with a table similar to Table 4 above. This continues until subject is indifferent with respect to the two situations.

In choosing a fixed state assessment procedure we are free to select

(1) A response format
  a. ends-in
  b. direct specification
(2) A comparison format
  a. standard fixed state
  b. paired binary gambles
  c. regional coherence
  d. local coherence
(3) Overall coherence checking by least squares
  a. yes
  b. no

The temptation for a person trained in both psychology and statistics to undertake the experimental comparison using some subset of a 2 by 4 by 2 factorial design is overpowering. Indeed the tooling-up for this experiment has begun including a further comparison with the fixed probability method and an investigation of comparative bias for central and extreme values of θ. For a Bayesian statistician, yielding to this temptation leads to a compulsion to state a prior distribution. In the absence of a precisely stated model this is not possible, but it is possible to state some general beliefs. I shall now do this and also invite you to attend the Psychometric Society meetings in May of 1980 where I shall report on the results of these experiments.

First, I believe that SFS with overall LSQ coherence checking will prove to be good but not best. Subjects find it hard to make unaided adjustments. As a result, incoherence will remain high, but overall fits will be tolerably good ($p=.7$). However, we are working on improvements that could make this procedure more attractive. I believe that the ends-in format will be preferred over direct magnitude estimation and will reduce the anchoring effect ($p = .8$). This may not hold for very experienced assessors who may find it tedious.

I believe that PBG will be unpopular and ineffective unless we find some simplification ($p = .9$). At present it is difficult and fatigueing and responses tend to be less than carefully considered.

I believe that regional and local coherence will both be useful and both will largely eliminate anchoring and adjustment biases ($p = .8$). I believe the regional coherence will be preferred by inexperienced users ($p = .6$) and local coherence by experienced users ($p = .7$, but perhaps only professional statisticians). Local Coherence provides a display of the large effect on extreme comparisons of minor adjustment in non-extreme comparisons. It is a powerful tool for locating the most desirable point in the probability range p^{+} .025. It is unclear to me whether overall least squares overfitting will be useful in conjunction with the RC or LC procedures, but my prior probability is .6 that it is.

Finally, for certain points I have high personal probability. Utility elicitation procedures can currently be conducted with accuracy and ease on CADA. Indeed, they can be conducted with sufficient ease and potentially with sufficient freedom from bias as to make applications of utility theory to education entirely feasible when the assessor is confronted with a specific problem of interest and importance, and when that problem is presented clearly and unambiguously.

ACKNOWLEDGEMENT

Research supported by the Office of Naval Research under contract N00014-77-C-0428.

REFERENCES

DEGROOT, M.H. (1970). *Optimal statistical decisions.* New York: McGraw Hill.

— (1975). *Probability and statistics.* Reading, Mass: Addison-Wesley.

FISCHHOFF, B., SLOVIC, P., & LICHTENSTEIN, S. (1977). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Tech. Rep.* PTR--1042-77-8, University of Iowa.

— (1980). Knowing what you want: Measuring labile values. Appearing in *Cognitive Processes in Choice and Decision Behavior.* (T. Wallston, ed.) Hillsdale, New Jersey: Laurence Erlbaum Associates, (In press).

GROSS, A.L., & SU, W.H. (1975). Defining a "fair" or "unbiased" selection model: A question of utilities. *J. Appl. Psychol.* **60**, 345-351.

HOGARTH, R.M. (1975). Cognitive processes and the assessment of subjective probability distributions. *J. Amer. Statist. Assoc.* **20**, 271-289.

KAHNEMAN, D., & TVERSKY, A. (1978). Prospect theory: An analysis of decision under risk. *Econometrica.* **47**, 263-291.

— (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology,* 3, 430-454.

KEENEY, R.L., & RAIFFA, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs.* New York: Wiley.

KNEPPRETH, N.P., GUSTAFSON, D.H. LEIFER, R.P. & JOHNSON, E.M. (1974). Techniques for the assessment of worth. *U.S. Army research institute for the behavioral and social sciences. Tech. Rep.* **254**, 40-53.

LEHMANN, E.L. (1959). *Testing Statistical Hypothesis.* New York: Wiley.

MOSTELLER, F., & NOGEE, P. (1951). An experimental measurement of utility. *Political Economy,* **59**, 371-404.

NOVICK, M.R. (1973). High School attainment: an example of a computer-assisted Bayesian approach to data analisis. *Internat. Statist. Rev.* **41**, 269-271.

— (1975). A cource in Bayesian Statistics. *Amer. Statist.* **29** , 94-97.

NOVICK, M.R. & ISAACS, G.L. (1978). Manual for the computer-assisted data analysis (CADA) Monitor, Iowa City: The University of Iowa.

NOVICK, M.R., ISAACS, G.L., HAMER, R., CHEN, J., CHUANG, D., WOODWORTH, G., MOLENAAR, I., LEWIS, C. & LIBBY, D. (1980). *Manual for The Computer-assisted data analysis monitor,* Iowa City: The University of Iowa

NOVICK, M.R. & JACKSON, P.H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw Hill.

NOVICK, M.R., & LINDLEY, D.V. (1979). Fixed state assessment of utility functions. *J. Amer. Statist. Assoc.* **74**, 306-311.

— (1978). The use of more realistic utility functions in educational applications. *J. Educational Measurement,* 181-192.

NOVICK, M.R. & PETERSEN, N.S. (1976). Towards equalizing educational and employment opportunity. *J. Educational Measurement,* **13**, 77-88.

PETERSEN, N.S. (1976). An expected utility model for "optimal" selection. *J. Educational Statistics.* **1**, 333-358.

PETERSEN, N.S. & NOVICK, M.R. (1976). An evaluation of some models for culture-fair selection. *J. Educational Measurement,* **13**, 3-31.

RAIFFA, H., & SCHLAIFER, R. (1961). *Applied statistical decision theory.* Boston: Harvard University.

SCHLAIFER, R. (1959). *Probability and statistics for business decisions: An introduction to managerial economics under uncertainty.* New York: McGraw Hill.

SLOVIC, P. (1975). Choice between equally valued alternatives. *Experimental Psychol.* **3**, 280-287.

— (1972). From Shakespeare to Simon: Speculation -and some evidence- about man's ability to process information. *Oregon Research Institute Research Bulletin,* **12**, 1-29.

SLOVIC, P., FISCHHOFF, B., & LICHTENSTEIN S. (1977). Behavioral decision theory. *Annual Rev. Psychol.*, **28**, 1-39.

SPETZLER, C.S. & STAEL VON HOLSTEIN, C.A.S. (1975). Probability encoding decision analysis. *Management Science*, **22**, 340-358.

SUPPES, P., & WALSH, K. (1959). A non-linear model for the experimental measurement of utility. *Behavioral Science*, **4**, 204-211.

SWALM, R.O. (1966). Utility theory--insights into risk taking. *Harvard Business Rev.* **44**, 123-136

TVERSKY, A. (1977). On the elicitation of preferences: Descriptive and prescriptive considerations. In *Conflicting objectives in decisions*, (D.E. Bell, R.L. Keeney and H. Raiffa eds.) 209-222. New York: Wiley.

TVERSKY, A. & KAHNEMAN, D. (1974). Judgment under uncertainty: Heuristics and biases *Science*, **185**, 1124-1131.

VAN DER LINDEN, W.J. & MELLENBERGH, G.J. (1977). Optimal cutting scores using a linear loss function. *Appl. Psychol. Measurement*, **1**, 593-599.

VLECK, C.A.J. (1973). Coherence of human judgment in a limited probabilistic environment. *Organizational Behavior and Human Performance*, **9**, 460-481.

VON NEUMANN, J. & MORGENSTERN, O. (1953). *Theory of games and economic behavior.* Princeton: University Press.

DISCUSSION

J.M. DICKEY (*University College of Wales Aberystwyth*):

I find the paper by Dr. Leonard stimulating. Many of us would agree with the statement that there is much more to real inferences than is modeled by Bayes' theorem: for example, that a given subjective-probability distribution might be usefully conditioned on new data at a particular time, but that to continue formally updating it to a sequence of new data over a long period without rethinking the probability model would be foolish. It would even be foolish to rely on Bayes' theorem on a single occasion if one closed one's mind regarding the assumptionsused.

But, of course, it is not necessary to close ones mind, nor ones eyes and ears. Bayesian theory does not require that, although it may seem so to some authors because of the silence in Bayesian theory on the subject of how to think up new models. The implications of coherence for the subject of learning from data have to do with what attitudes to take regarding contingent bets, how to reason now about the information in future data. The axioms of coherent potential behaviour do not imply that, after the data is in, one should actually follow the previous plan in updating ones opinions. That is, probability conditioning (for example, Bayes' theorem) is not necessary for real opinions, but it provides a point of reference, a rational yardstick, a standard relationship between prior and posterior opinions. If ones opinions do not obey probability conditioning, then one looks for a reasonable probability model under which they do, or which implies opinions that one can reasonably adopt.

How should a Bayesian statistician look at his data to see **whether** he will need to think up new models? Karl Popper (1972) imagines scientific research as a continuing process of using experimental data to test the validity of theories which are then revised

when rejected by tests. Popper's nonBayesian conception also suffers from silence on the subject of how to think up new models. Also, it inherits a defect from traditional statistical "data analysis" on the subject of how to decide whether a new model is needed. This defect in traditional tests for validity-checking of models was pointed out by Berkson (1938). In practice, no model ever tested is exactly true, and any prespecified model will be rejected for a large enough (fixed-size) sample. This makes acceptance of models largely a question of the size of samples taken. (See Kadane and Dickey, 1979, for a Bayesian discussion of this problem). Another question, for which a traditional statistician's answers can only be highly subjective when no alternative models have been suggested, is the more general question of which validity tests to perform. Which *experiment* to perform also remains largely subjective.

So traditional theory and Bayesian theory are both limited in the scope of their application. I think it is a mistake, though, to say that coherence implies complexity or that coherence misleads. Do the rules of logic or arithmetic mislead? Nor does IMP to my mind "oppose" coherence, unless Dr. Leonard insists on tying IMP to the Freudian notion of Id. I agree that IMP seems complex, but I call on Dr. Leonard and others to develop theory to shed light on its mysteries.

Dr. Leonard reminds us of the old question of discrimination methods versus regression analysis. It is really simpler for the statistician to specify $p\ (\mathbf{x}|\Lambda)$ than $p\ (y|\mathbf{x})$? I note that he suggest the use of estimated sampling probabilities to approximate predictive probabilities, while Aitchison and Dunsmore (1975) recommend the use of predictive probabilities to estimate sampling probabilities.

Finally, it is claimed that the Bayes factor is sensitive to the choice of conditional prior density, and increasingly so for increasing sample size. Of course, in practice the Bayes factor goes to zero or to infinity as sample size increases. A very small or very large Bayes factor is strong evidence for or against the more complicated model, respectively. So it remains to be shown that the "sensitivity" happens before the evidence becomes too strong to be refuted by the changes in the Bayes factor wrought by reasonable perturbations in the prior density.

My comments on Professor Novick's paper joint with Dekeyrel and Chuang would seen to apply with equal force had the paper been concerned with probability assessments, rather than utility assessments. (Utilities are equivalent to probabilities in technical senses, and this equivalence is exploited in their assessment methods). Therefore, I should like the authors to consider my comments with an eye to the possibility that I have failed to appreciate properties inherent only to utilities. Perhaps they would bring out the important differences in their reply to this discussion.

The methods given are ingenious and rather elegant. A person wishing to use them to assess his own utilities would, I feel sure, need to spend appreciable time and effort learning to use them as effective tools. The worry, of course, is that in so doing the person may acquire bad habits or "biases" that would connect up his different uses of the tool, rather than connecting together the tool and his underlying utilities.

Instead of a "person", the authors refer to a "subject". This latter term has been reserved in the psychological literature to mean the same as "object", in the spirit of conceiving persons other than oneself as machines. One trouble with this conception is that it just does not work well, except at a mere physiological level. Persons do not behave predictably without reference to context, including the histories of their personal attitudes and social settings (Kelly, 1955). Experiments tend to be aimed at discovering simple universal context-free laws of behavior, such as, laws that would favour this assessment tool over that one. What is it that justifies our thinking that isolated laboratory experiments will yield psychological findings of any importance in real-world applications?

In spite of the doubts expressed here, I should like to urge the authors to carry out the experiments envisaged, preferably in real applications.

W.H. DUMOUCHEL (*Massachusetts Institute of Technology*):

Professor Leonard's emphasis on the necessity to develop workable procedures, and to show our colleagues that they do work, is well put, in my opinion. More focus is needed on what we can do, rather than too much concentration on the logical inconsistencies of classical statistics. Strict consistency is often unattainable in the real world. For example, we all know that prior distributions cannot logically depend on the data. Yet Professor Leonard rightly points out that most responsible statisticians, Bayesian or not, will try to obtain a "feel" for the data with plots, etc., before inducing a likelihood function or even deciding on a parameter space. However, I am not so pessimistic as to rule out a useful Bayesian approach to many "global" problems. Often a mixture of two or three models can quite well capture the essentials of even a fairly complicated situation, and thus help derive real-life conclusions from the data. The binomial example of section 4 does not seem convincing to me. The situation is that of choosing between H_0 and H_1 based on the observations of n exchangeable observations of 0 or 1, whose sum is x

$$\begin{aligned}&H_0 : x \sim P_0(x)\\ &H_1 : x| \sim Bin(n,\theta)\\ &\qquad x = 0,1,\ldots,n \qquad\qquad \theta \sim \text{Beta}\,(\alpha,\beta)\end{aligned}$$

The supposed paradox is that posterior odds ratio of H_0 *vs* H_1 depends importantly on α and β even as $n \to \infty$, especially if x/n is far from $\alpha/(\alpha+\beta)$. But the fact of n being large here does not reasonably imply that the sample information should "swamp" the prior information. When alternative H_1 is true and n is large, the variation in $\theta = x/n$ is negligible conditional on θ, so that the relevant comparison is

$$\begin{aligned}&H_0 : \theta \sim P_0(n\theta)\\ &H_1 : \theta \sim \text{Beta}\,(\alpha,\beta)\end{aligned}$$

Thus the problem is more like that of deciding whether a single observation could have a particular beta distribution, and naturally the parameters of that beta distribution would play an important role in the decision.

On another point, in spite of my own liking for logit probability models, I suspect they are being oversold in section 7. The author's distinction between a probabilistic and a predictive model eludes me. Two possible interpretations are: (1) the full information versus conditional information approach to contingency tables, or (2) the errors in variables problem of regression. But the further discussion doesn't seem relevant to either interpretation. The author seems to imply that multivariate density estimation is simpler and more reliable than more common procedures such as stepwise regression. I would guess that use of one of the various robust regression techniques now widely available would be more fruitful than abandoning the ordinal structure of the response variable in favor of a purely categorical-data approach.

Finally, as an argument against constructing unnecessarily complicated models, the author states in section 8 that modeling a thick-tailed distribution is unnecessary if, even with a normal model, the real-life conclusions are the same with and without inclusion of the outliers in the analysis. This cannot be true in general, as the following example shows. Suppose that a sample of size $n = 100$ has mean 0 and standard deviation 1, with one or two outliers near the value $x = 4$. Suppose further that the real-life problem is to decide whether Prob $(X>4)<.001$. Then a normal model including the outliers would estimate Prob $(X>4)<10^{-4}$, while excluding the outliers would result in a smaller sample standard deviation and an even smaller estimate for Prob$(X>4)$. Yet fitting the data to most families of thicktailed distributions would estimate Prob$(X>4)$ to be near the sample proportion, namely 0.01.

Professor Novick and his co-authors are to be commended for continuing to explore a topic so vital to the practical functioning of the Bayesian method. Until we can show how prior opinion can be elicited in a workable fashion, the subjective Bayesian viewpoint can hardly proliferate. The present paper considers with care and sophistication a simple problem involving a single, ordered attribute, and makes us very conscious of how much harder a more realistic elicitation involving several dimensions and a complex data set would be. The work of Kadane *et.al.* (1979) combined with the present paper provide a start toward computerizing this process.

The author's references to the work of Amos Tversky and his associates are welcome. Certainty bias and anchoring bias are present not only in elicitation problems, and overcoming them can be used as a theme for data analysis in general. Whenever we tell our elementary statistics classes to be more conscious of variation, we are fighting the certainty bias, and when we teach proper methods of estimation we counter the anchoring bias. But Bayesian methodology is peculiarly affected, on a second level, by these tendencies. A stronger potential barrier to solution of the elicitation problem is raised by the work of Shafer (1976) who argues that human opinions are too complicated to be represented by simple probability distributions or utility functions. I would be interested to know if experiments such as the present authors are performing could be designed to test this or similar propositions.

In any event, the "local", "regional", and "ends in" procedures presented here seem reasonable and clever and I am looking forward to the results of the authors' future experiments. There are just a few more specific questions that come to mind:

a) What if the ordering of the states is not prescribed? Would your methods change?

b) Although elicitation of probabilities is formally identical with the elicitation of utilities, the psychological reactions of subjects may differ for the two tasks. Is there any evidence of this?

c) What evidence is there that the regression on the log odds scale is optimal for the coherence checking algorithm? Might some weighted regression be better? Is the standard error of the residuals a useful number?

c) How much real time do these elicitations take? How long for a novice to elicit all the factors for the probabilities in a 2x2 table, and what fraction of them show noticable fatigue and/or boredom before finishing?

I hope that these questions will help stimulate the authors to continue their interesting work.

J.M. BERNARDO (*University of Valencia*):

I certainly believe that the idea used by Professor Novick of requiring the decision maker to give more than the minimum number of judgments in fitting a personal probability distribution or utility function is important and very useful. I wonder however what is the coherent justification for using least squares in order to force coherence among those judgments

S. FRENCH (*University of Manchester*):

Firstly, perhaps Dr. Leonard will forgive my pointing to an unfortunate omission in his paper. In quoting DeGroot's axiom system for subjective probability, he omits the CP axiom (DeGroot (1970), Chapter 6). It is the CP axiom that introduces the notion of conditional probability and hence justifies the use of Bayes Theorem. Without the CP axiom this system does not pretend to justify Bayesian inference. If Dr. Leonard wishes to criticise the use of axiom systems, he really should cite a whole system.

Turning now to the paper of Novick, Dekeyrel and Chuang, I have two questions that I should like to ask. First, in the fixed state method of assessment the values $U(\theta_o) = 0$, $U(\theta_N) = 1$ are fixed. The values of $U(\theta_n)$ for intermediate n are determined by relations of the form

$$U(\theta_n) = p_n U(\theta_{n+1}) + (1\text{-}p_n) U(\theta_{n-1}) \qquad (*)$$

Now, since the paper's very essence is to admit incoherence on the part of the decision maker's statements, it must be admitted that the p_n are "in error". Does this error transmit itself evenly to the determination of $U(\theta_n)$ or does the error on the $U(\theta_n)$ rise steadily from 0 on $U(\theta_o)$ to a maximum on $U(\theta_{N/2})$ before falling away to 0 again on $U(\theta_N)$? There is a relevant passage in Spetzler (1968) in which he discusses the relative merits of three different methods of measuring utility.

My second question concerns the decision makers' role in the resolution of incoherence. For me one of the basic aims of decision analysis is to bring understanding. In particular the process of introspection is not simply one of measuring utilities and subjective probabilities. Rather it is a process that helps the decision maker explore his preference belief structure, discover inconsistencies, think about them and then resolve them. It seems imperative to me that of method of construnting a decision maker's utility function should always refer back to him any discovered inconsistency so that he may reconsider his preferences. Only when all the inconsistencies are of such a slight nature that it is beyond the decision maker's powers of discrimination to resclve them, should an automatic resolution process be invoked. Do I understand that the authors' procedure does in fact do this, namely only use least squares with coherence constraints as a tidying up device having left all the major resoution of inconsistency to the decision maker?

J.B. KADANE (*Carnegie-Mellon University*):

In the discussion, both Dennis Lindley and Bruce Hill strongly criticized Tom Leonard's paper for not being sufficiently Bayesian. In doing so, I think that they have overreacted. When a Bayesian does statistical modelling and data analysis, compromises are often necessary to keep control of the analysis, to separate what is important from what is not.

To associate Tom Leonard's position in this paper with Glenn Shafer's, as did Bruce Hill, is to mix two very different positions, I think. As I understand Shafer's ideas, he rejects Bayes Theorem and the Bayesian paradigm as a theory. This seems to me very different from Leonard's position, which keeps Bayesian theory as essential background for doing statistics. To associate these positions does an injustice to both Leonard's and Shafer's positions.

D.V. LINDLEY (*University College London*):

I find myself in almost total disagreement with the views expressed in Leonard's paper. Coherence becomes more important the bigger the situation, not less. If only one uncertain event is assessed, then coherence does nothing more than assert that the descriptive number lies between 0 and 1. With two events, A and B, coherence begins to play a more important role: for example, $p(AB) = p(A)p(B|A)$. The more events, the more opportunity there is to exploit coherence and the more necessary it becomes to do so.

Perhaps it is this fallacious view that leads to Leonard attaching importance to Axiom 5. All this axiom does is to tie probability to a numbering system: the multiplication and addition rules, the rules of coherence, are really contained in the earlier, important axioms and his omitted axiom of called-off bets. Probability is not just a number between 0 and 1: it is a number obeying two important rules of combination.

A. O'HAGAN, (*University of Warwick*):

Dr. Leonard's *IMPs* are of course an over-complication, providing no real insight into the processes of practical statistics. But if we regard them as merely a thin excuse for presenting a miscellany of ideas - his sections 4 to 12 then there is much food for thought in his paper. I would like to examine just a few of the snapshots in Dr. Leonard's album.

His skewed-normal distribution of section 8 is ingenious, but I wonder if some of his criteria (i) to (vii) were chosen *a posteriori*. I commend to him the skew distribution derived in O'Hagan and Leonard (1976), for which I think we could draw up an equally impressive list of criteria. For instance it is more tractable than the skewed-normal.

The sensitivity of the Bayes factor (4.3) to the prior hyperparameter a in his binomial example of section 4 could be quite worrying. Some insight is obtained initially by ignoring θ. Since $D(\alpha,\beta)$ is simply the prior (marginal) probability of the frequency x under the binomial model, we are just comparing the two simple hypotheses given by the distributions $p_0(x)$ and $p_1(x) = D(\alpha,\beta)$. The observed value of x discriminates strongly between the hypotheses if the ratio $R_. = p_0(x)/p_1(x)$ is very large or very small. Dr. Leonard introduces a third hypothesis, that x has distribution $p_2(x) = D(\alpha + 1,\beta)$ and observes that it may be possible to find an x which does not discriminate strongly between p_0 and p_1 but does discriminate strongly between p_0 and p_2. He does this by showing that the ratio (4.5) can give an x that discriminates strongly between p_1 and p_2. His thesis is that this odd because p_1 and p_2 are very similar. But with most parametric families of distributions we can find observations discriminating strongly between any two members of the family, however close their parameter values may be. Consider for example the distributions $N(0,1)$ and $N(\epsilon, 1)$: however small $|\epsilon| > 0$ is, as x tends to infinity the likelihood ratio

$$\exp\{-\tfrac{1}{2}x^2 + \tfrac{1}{2}(x-\epsilon)^2\} = \exp(-x\epsilon + \tfrac{1}{2}\epsilon^2)$$

tends either to zero or to infinity. Almost all the parametric families in common use have monotone likelihood ratios (see Lehmann (1959)) and in most cases the likelihood ratio is unbounded. In fact, since Dr. Leonard's beta-binomial has a bounded likelihood ratio (for given sample size), he has chosen one of the less convincing examples of "sensitivity". Other examples may be constructed similarly "— $p_1(x)$"— is formed from a prior distribution for a scalar parameter θ indexed by a prior hyperparameter ϕ, and a sampling distribution for x given θ. Whenever these two distributions have monotone likelihood ratios, e.g. any two exponential-family distributions (Lehmann, p. 70), then $p_1(x)$ will have a monotone likelihood ratio in ϕ (Lehmann, p. 343 problem 7).

Therefore, Dr. Leonard's sensitivity problem arises whenever we deal only in exponential families. Having seen the "problem" in the above terms I feel that it is not as unreasonable as he implies, but I do think that it is important to recognise that nearly all commonly used distributions will lead to this kind of behaviour and that radically different behaviour is possible using distributions with non-monotone likelihood ratios. In O'Hagan (1979), and more explicitly in a follow-up paper submitted to the Annals of Statistics, I have made this point in connection with a different kind of behaviour which always results from using distributions with monotone likelihood ratios, and not otherwise. In his section 9, Dr. Leonard criticises exponential families on even more fundamental grounds. It is time that we looked very seriously beyond the convenient, tractable exponential families because they are severely limiting the kinds of inference that we can make.

A.F.M. SMITH *(University of Nottingham)*:

Leonard seems to be making two rather strong attacks on the axioms. If I understand him correctly, he states that:

(i) the straightforward claims set out in 2*a*) and 2*b*) are much more directly *compelling* to clients than are the axioms; and, in any case, they are more *honest*;

(ii) the axioms are tautologous.

Let us first consider (i), and recall that statement 2*a*) invokes the phrases "much more reasonable", while statement 2*b*) refers to "superior practical results". Does Tom Leonard really believe that these particular phrases can (honestly) command general acceptance as having directly obvious meanings that require no further analysis? And if someone refuses to accept these as primitive terms of reference, I think I know where Tom Leonard would eventually end up in attempting an unambiguous explication of "reasonable" and "superior" - back at this axiom system!.

The criticism in (ii) seems most peculiar!. *Theorems* deduced from the axioms are, *of course*, "contained in" them in the sense Tom Leonard presumably intends. But, surely, the (for us) rather profound methodological implications - the likelihood principle, the need to integrate out nuisance parameters - are *in no way* obviously "contained in" the axioms in the sense that they are directly intuited (or guessed, even) by someone who contemplates the axioms?

T.W.F. STROUD *(Queen's University Canada)*:

Leonard's article presents a refreshing relief from doctrinaire approaches which begin with a statement of the statistician's model and his prior beliefs about the parameters of the model. In fact, the statistician always has to begin with a real-life process and, hence, any model concerning this process (and, consequently, any prior distribution on the parameters of such a model) must be regarded as very tentative.

Sections 4 and 5 focus on some important facts often overlooked by Bayesian statisticians. In Section 4 it is pointed out that probabilities associated with choosing *between* models may be quite sensitive to the choice of prior distributions *within* models. Because inference *within* a model is insensitive to prior information when samples are large, it is easy to think that in large samples the prior doesn't matter. But the thing which makes the prior not matter is the likelihood, which is completely model-based. The example presented in Section 4 shows that, in situations where the prior mean within the binomial model ξ is very different from the sample mean p, the information in the data which is ancillary to the binomial model (which is what we need for testing the model) may *not* swamp out the prior in moderately large samples.

In Section 10, which deals with problems involving hyperparameters, the method of maximizing the marginal likelihood is advocated as an alternative to specifying "complicated and possibly confusing" prior distributions on the hyperparameters. Whereas in many problems maximizing the marginal likelihood gives virtually the same answer as integrating over a locally uninformative prior on the hyperparameters, no justification has been given that the former procedure is anything but a convenient approximation to the latter. In some cases, the approximation may be poor. For example, in the normal one-way classification shrunken estimates of the group means toward the grand mean may be obtained by putting a conjugate prior on the exchangeable group means and estimating the hyperparameters in this prior by maximum likelihood (Stroud, 1980). But if the number of groups is small (say 3 or 4), this procedure shrinks too much toward the grand mean because the likelihood function of the between-within variance ratio is skewed, causing the mode to underestimate this variance ratio. A similar problem exists if one uses a prior on the hyperparameters but then resorts to substituting the posterior modal values of hyperparameters, rather than integrating over them. In such cases where skewness causes a problem one should either integrate out the hyperparameters or devise a technique for suitably adjusting the modal estimates in the direction of the skewness.

REPLY TO THE DISCUSSION

T. LEONARD *(University of Warwick)*:

Many thanks to the discussants for their helpful contributions which seem to provide a good representation of current Bayesian thought about the area of Statistics. Since the conference Dennis Lindley and I have corresponded in detail about the axioms, and this has helped us to clarity our ideas in this area.

A positive contribution of this correspondence was an indication that my Axiom *5a* is not needed in the very strictest mathematical sense, as De Groot utilizes the mathematical properties of random variables to their fullest extent (they are A-measurable functions from the parameter space to the real line). However, if the outcomes of the auxiliary experiment were simply regarded as numerical values, then my Axiom *5a* would be needed to link the auxiliary experiment with the parameter space: it is this interpretation which the probability assessor would utilize when actually carrying out the suggested procedure. Moreover, my axioms 5 and *5a* are equivalent mathematically to the combination of De Groot's Axiom 5, and his assumption of A-measurability of the random variable. Therefore my comments are relevant whichever interpretation is used; it is my firm understanding that the combination of the first four axioms with the assumptions surrounding the fifth axiom should be viewed in an inductive sense as virtually as strong as the final result. I would however like to thank Dennis for indicating the desirability of clarification of this mathematical point.

It still seems completely obvious to me that the axioms are not really proving much, but simply describing a way of thinking. During my correspondence with Dennis he suggested various sensible changes to the axioms, but despite about half-a-dozen intuitively appealing suggestions at least one of the axioms always turned out upon close scrutiny to be similar in strength to De Groot's fifth axiom. It is interesting that whilst recently teaching utility theory, I decided to play the role of a formal Bayesian, but this approach was quickly shown to be deficient by a series of simple and unprompted questions from my students; these were much on the same lines as the points I have raised here about subjective probability.

Dennis seems to have dodged the real issue - my main point is that coherence is less important and even constrictive in practical situations where the objective is to extract real-life conclusions from a data set. Probably we Bayesians should leave our ivory towers once in a while and work in a Statistical Laboratory analyzing real data. We might then learn that modelling is the really important part of statistics; analyses which proceed conditionally upon the choice of model are enjoyable but do not provide the complete answer.

I would like to thank Tony O'Hagan for his comments. I don't think that my *IMP*'s are an over-theoretisation - in fact there're not really a theoretisation at all! They are just a way of thinking, or perhaps a term to describe what most of us have been doing anyway. My point is that thinking about the problem in order to extract a model or a conclusion is much more important than trying to be formally coherent. Tony's comments on the sensitivity problem are helpful and interesting. His work on outlier behaviour would be useful if it were possible to find families of distributions with thick tails which are both meaningful and analytically tractable, for example, in multivariate situations.

I'm a bit confused by Simon French's comments. I didn't use the conditional probability axiom because I was just discussing straight-forward probability. I think however that my main points would extend to this situation.

Tom Stroud's thinking seems to be on similar lines to my own - we should probably form a clique of pragmatic Bayesians (this may be a good time to announce the foundation of the Bayesian-Fisherian school of statistics!). It is possible to justify estimating hyperparameters by their marginal likelihood estimates when the number of first-stage parameters is greater than about ten, because the estimates will then approximate the Bayes estimates under a wide range of loss functions. When the dimensions are smaller the estimates are less precise but still fairly sensible. A more sophisticated estimation procedure would in this case probably not be justified in view of the small amount of information available about the hyperparameters.

Bill DuMouchel's comments are very helpful and I'm glad that he supports the main theme of my paper. I remain a bit pessimistic about a mixed model approach since it would not be particularly meaningful or easy to check out each of the candidate models against the data or to think in a lucid way about the complicated analysis employed. It is interesting that he indicates that the binomial hypothesis testing problem is similar to deciding whether a single observation could have a particular beta distribution - this really supports my argument since it tells us that the standard Bayesian procedure for this situation can't properly distinguish between the two hypotheses.

My distinction between probabilistic and predictive models is a practical one. For many data sets the explanatory variables are extremely noisy so that it is virtually impossible to find a least squares model via standard procedures like stepwise regression, and therefore difficult to get reasonable numerical predictions of further dependent variables. However the data may still be rich in a content of a probabilistic nature, in the sense that they indicate how much the statistician should adjust his probabilities about the dependent variables, in the light of knowledge of the explanatory variables. In such circumstances, where we just can't find a reasonable least squares model, we can often still arrive at useful conclusions by modelling the distributions of the important explanatory variables.

I am not arguing completely against the use of thick-tailed distributions, but simply saying that if we look at the data and think about the problem then we can sometimes avoid this extra complication. In the example Bill discusses, I guess that most of us would prefer a much smaller value for prob($X>4$) than 0.01.

Adrian Smith feels that my implication that the axioms of coherence are tautologous is most peculiar. This is probably because, like Dennis, he is thinking deductively rather than inductively - if we constrain ourselves to Bayesian formalism then statements by more open and inductive thinkers will very often appear to be peculiar. As I see it, if we look at the axioms and judge intuitively the strength of what is being assumed, and next look inductively at the strength of the final result, then the two appraisals will be extremely similar. Therefore the fact that the axioms deductively imply the final result does not really give us much - it would be inductively speaking just as reasonable to assume the final result to start off with. It's a pity that neither Dennis nor Adrian have taken this opportunity to look deeply enough at the problem to be able to give a definitive answer to this point.

I can't see how the likelihood principle follows from the axioms unless coherence is also assumed across an n-dimensional sample space in order to justify the existence of a sampling distribution - an extremely complicated assumption (don't the sufficiency principle and the very complex conditionality principle come into it as well?). The assumption that we can marginalise subjective distributions is barely stronger than the axioms that might be used to justify this procedure.

Further analyses are of course needed to justify statements like "superior practical results", but I think that this has already been done - see for example the work by Adrian and others on multi-parameter estimation, time series analysis, and categorical data. I personally think that the Bayesian approach is "much more reasonable" because it is extremely natural to think in terms of probability distributions when updating information about quantities of interest.

My thanks to Jim Dickey and Jay Kadane for their contributions. On the question of discrimination methods versus regression analysis it is indeed much simpler in many situations to model the distributions of the explanatory variables. Of course, one should always choose the method which best suits the practical situation at hand.

I would finally like to say how much I enjoyed giving a paper in the same session as Mel Novick. His practical implementation on CADA of my early marginalization work on categorical data fits in well with the things I have been trying to say.

M.R. NOVICK *(University of Iowa)*:

The commentary provided by Professors Bernardo, Dickey, Du Mouchel and French, are useful in themselves, but to me they have the added value of opening up for discussion some topics that I might have covered in my original presentation, had time and foresight permitted.

Professor Bernardo notes, with bated foil, that there may be no "coherent justifications for using least-squares in order to force coherence among ... judgments". He is, of course, correct. The only reply is that coherence, like virtue, can be absolute only in contemplation and is more likely to be compelling as we examine the actions of others rather than ourselves. Wisdom must guide us in knowing when small deficiencies in coherence (and virtue) can be tolerated.

The essence of Professor Dickey's critique of our paper is summarized in his question: "What is it that justifies our thinking that isolated laboratory experiments will yield findings of any importance in real-world applications?" Feelings of inadequacy in my ability to contribute anything *new* to the discussion of *that* question compel me to refer Professor Dickey to his biologist, chemist, physicist, psychologist, et. al. friends, some of whom may be willing to take the time to instruct him on the general decline in acceptance of the Kantian view of science and the acceptance since the end of the Dark Ages of the value of laboratory experimentation. For my own part I shall borrow Professor Bernardo's bated foil and ask Professor Dickey, "What is it that justifies *his* thinking that the mathematical derivations *he* presents us without *any* empirical investigation of relevance, will provide us with useful methods of assessing prior probabilities?" Perhaps Professor Dickey and I are both guilty of demanding a higher level of virtue and coherence of others than of ourselves. For my part I speculate that Professor Dickey's work will be very useful but question the appropriateness of his presupposition.

Professor Dickey, however, is not entirely off the mark. We have found that our methods are "successful" only when we go to great lengths, in our laboratory, to simulate practical decision problems. People do not carry around utility functions in their heads and we ought not to view the assessment process simply as a psychological measurement (psychometric) problem. However, we have also found that the nature of the graphic display has significant influence on assessors responses and that the anchoring effect can be reduced by the methods we propose. We also believe that further refinements will be useful.

Professor Du Mouchel's comments are more penetrating and require more detailed response. It is true that human opinions can be very complicated. Part of that complication is due to incoherence which, it is hoped, can be reduced through computer interaction. It is also true that humans attempt to uncomplicate their opinions and decision processes by the use of simplifying heuristics. Unfortunately these heuristics *typically* introduce bias. Our goal is to uncomplicate human opinion by providing alternative heuristics that avoid major biasing effects. This is not a simple task and we make no claim of "complete" success. But if, in education, I had to choose between decision-making with or without the prior probability, utility assessment, and decision-making procedures now available on the Computer-Assisted Data Analysis (CADA) Monitor I would certainly opt to use CADA.

With respect to Professor Du Mouchel's question as to whether experiments could be designed to test whether human opinions are too complicated to be represented by simple probability distributions or utility functions, I would respond that I think rather different experiments are necessary. I personally accept the notion that human opinion is too complicated to be so modelled. The point, however, is that what we seek is *not* a descriptive modelling of what human opinion *is*, but a normative modelling of what a particular human being's opinion "ought" to be. The word "ought" here has a special meaning that must be made precise. A human being's opinion "ought" to be internally coherent and ought to be consistent with contemplated behavior. If contemplated behavior is inconsistent no formal modelling with a probability distribution or utility function is possible. Thus probability and utility assessment procedures do not involve descriptive modelling. They involve a process that changes opinions in some way that results in internal coherence without changing those aspects of contemplated behavior that most clearly represent the person's opinions regarding the real world.

I now respond to Professor Du Mouchel's specific questions a) to d):

a) If states are not ordered we begin by ordering them.

b) All of our elicitation procedures require probability judgements (fixed state as opposed to fixed probability). We believe that the direct elicitation of utilities is deceptively easy but subject to a high degree of artifactual bias.

c) We have our intuition and some informed observation to suggest benefit from the log-odds scale for the regression of probabilities. I have very high personal probability that this is very much better than least-squares in the original metric. However, I would think that somewhat less weight on the extreme values might be useful. Dennis, Lindley and I have often debated the relative benefits of log-odds and root inverse sine transformations.

d) For most problems that we have adressed to date elicitations are handled quickly, with perhaps 10% of subjects showing boredom, fatigue, or uncorrectable incoherence. (For some this result may be endemic to the laboratory context which remains somewhat artificial despite our best efforts). The key to success with such methods is the moderate realism of the established scenario and the smoothness of the person/machine interaction. But our degree of success does also vary with the complexity of the model. A nine point unidimensional utility assessment is comfortable. A bivariate utility assessment is more difficult. Higher dimensional assessment is currently beyond our ability. (We have not been impressed by the mathematically convenient but largely unrealistic assumptions that others have chosen to make). The interrogation procedure for multiple linear regression originally programmed following the Kadane et. al. suggestions proved inadequate. However, Dr. James Chen of my staff has now produced an acceptable program which is tolerated by keen investigators, but is still wearisome for most users. Further improvements will need to be made.

Finally, let me adress Professor French's useful queries. Professor Lindley and I showed in our original paper that the value of P_n effected $U(\theta_n)$ most with decreasing effect for more distant values of θ_i. This is, I think, a desirable property, though independence for $i \neq n$ would be preferable.

Professor French's second query gets to the heart of our methods and I am grateful to him for raising the issue because I neglected this vital point in my presentation. (I really ought not assume that everyone is familiar with our CADA project). If I may borrow Professor French's words, the primary function of elicitation procedures on CADA is to help the "decision maker explore his preference belief structure, discover inconsistencies, think about them and then resolve them" We believe that this process is facilitated by conversational language computer interaction. Descriptions of CADA are contained in my article on CADA in the *International Statistical Review*, 1973, my article in the *American Statistician* in 1975 and a second article in the *American Statistician* to appear in November, 1979.

REFERENCES IN THE DISCUSSION

AITCHISON, J. and DUNSMORE, I.R. (1975) *Statistical Prediction Analysis*. Cambridge: University Press.

BERKSON, J. (1938) Some difficulties of interpretation encountered in the application of the chi-squared test. *J. Amer. Statist. Assoc.* **33**, 526-42.

DEGROOT, M.H. (1970) *Optimal Statistical Decisions*, Reading, Mass: Addision-Wesley.

KADANE, J.B. and DICKEY, J.M. (1979) Bayesian decision theory and the simplification of models. *Evaluation of Economic Models*. (J. Kmenta and J. Ramsey, Eds.) New York: Academic Press.

KADANE, J.B., DICKEY, J.M., WINKLER, R.L., SMITH, and PETERS, S.C., 1979. *Tech. Rep.* **150**, Carnegie-Mellon.

KELLY, G.A. (1955) *The Psychology of Personal Constructs*. New York: Norton.

LEHMANN, E.L. (1959) *Testing Statistical Hypotheses*. New York: Wiley.

O'HAGAN, A (1979) On outlier rejection phenomena in Bayes inference. *J. Roy. Statist. Soc. B*. **41**, 358-367.

O'HAGAN, A. and LEONARD, T. (1976) Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* **63**, 201-203.

POPPER, K. (1972) *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.

SHAFER, G.: (1976) *A Mathematical Theory of Evidence*. Princenton: University Press.

SPETZLER, C.S. (1968) I.E.E.E. Trans. Systems, Science Cybernetics. SSC-4, 297-300.

STROUD, T.W.F. (1980) Empirical Bayes versions of Stein-type estimators. *Tech. Rep.* Stanford University.

Trab. Estadist. (1980)
Vol. 31 pp.

J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.
Bayesian Statistics (1980)
pp. 583-648 Valencia, Spain: University Press

14. Hypothesis Testing

INVITED PAPERS

ZELLNER, A. and SIOW, A. (*University of Chicago*)
Posterior odds ratio for selected regression hypothesis

BERNARDO, J.M. (*Universidad de Valencia*)
A Bayesian analysis of classical hypothesis testing

DISCUSSANTS

JAYNES, E.T. (*Washington University*)
SPIEGELHALTER, D.J., (*University of Nottingham*)
AKAIKE, H. (*Institute of Statistical Mathematics, Tokyo*)
DEMPSTER, A.P. (*Harvard University*)
DICKEY, J.M. (*University College Wales Aberystwyth*)
GEISSER, S. (*University of Minnesota*)
GOOD, I.J. (*Virginia Polytechnic and State University*)
LINDLEY, D.V. (*University College London*)
O'HAGAN, A. (*University of Warwick*)
ZELLNER, A. (*University of Chicago*)

REPLY TO THE DISCUSSION

Posterior Odds Ratios for Selected Regression Hypotheses

A. ZELLNER and A. SIOW
University of Chicago

SUMMARY

Bayesian posterior odds ratios for frequently encountered hypotheses about parameters of the normal linear multiple regression model are derived and discussed. For the particular prior distributions utilized, it is found that the posterior odds ratios can be well approximated by functions that are monotonic in usual sampling theory F statistics. Some implications of this finding and the relation of our work to the pioneering work of Jeffreys and others are considered. Tabulations of odds ratios are provided and discussed.

Keywords: BAYESIAN ODDS RATIOS; HYPOTHESIS TESTING; REGRESSION HYPOTHESES; REGRESSION MODEL.

1. INTRODUCTION

In this paper we derive posterior odds ratios for selected sharp hypotheses which are frequently encountered in regression analysis[1]. Our approach involves use of generalized forms of Jeffreys's prior distributions that he regards as appropriate when there is little previous information, that is "...in the early stages of a subject...," Jeffreys (1967, p. 252). Of course if more information is available, more informative prior distributions can of course be employed as has been done by Dickey (1971, 1975, 1977), Leamer (1978), Zellner (1971, p. 307 ff.) and others. Herein, we shall emphasize the situation in which little is known and, as will be seen resulting posterior odds can be expressed in terms of usual t or F statistics and degrees of freedom. Thus the

[1] See Jaynes (1976) for valuable analyses of a number of important practical examples illustrating the need for care in formulating relevant hypotheses and using appropriate techniques in order to obtain sensible results.

results to be presented provide a direct small-sample link between Bayesian posterior odds ratios and non-Bayesian test statistics as in the previous work of Jeffreys (1957, 1967, 1978), Lindley (1957) and others. Also, some large sample connections between Bayesian posterior odds ratios and non-Bayesian large sample test statistics are developed which are special cases of the general results of Lindley (1961) and Schwarz (1978).

Several, including Thornber (1966), Geisel (1970), Geisel and Gaver (1974), Leamer (1978), and Lempers (1971) have considered posterior odds ratios for regression hypotheses when little information is available. Our approach differs from those utilized in these works in that we employ prior distributions different from those employed in these works.

Since our approach is an extension of that originally presented by Jeffreys (1967, Ch.V.), we present a brief review of Jeffreys's related results in Section 2. In Section 3 posterior odds ratios for several important regression hypotheses are derived. Section 4 presents some numerical evaluations of the posterior odds ratios derived in Section 3 while a summary of results and some concluding remarks are given in Section 5.

2. REVIEW OF JEFFREYS'S RESULTS

Jeffreys (1967, Ch.V) has derived posterior odds ratios for a number of important testing problems in which little prior information is available and the issue is whether a parameter's value is equal to zero, a sharp null hypothesis. A sharp null hypothesis of "no effect" is frequently encountered and thus it is important to have an analysis of it. Jeffreys refers to such an analysis as "significance testing" and contrasts it with an estimation approach in which no special value of the parameter, for example zero is singled out for special attention. Also, he (1967, p. 251) points out that his estimation prior probability density function (pdf)-for representing "knowing little", for example a uniform prior *pdf* is inappropriate for a significance testing situation in which little is known about a parameter's value[2].

To be specific, consider Jeffreys's (1967, p. 268 ff.) analysis of the normal mean problem,

$$y_i = \lambda + u_i \qquad i = 1,2,\ldots,n \tag{2.1}$$

where the y_i's are observations and the u_i's are unobserved errors assumed independently drawn from a normal population with zero mean and standard

[2] In regression analysis when we delimit the number of regressors to be finite, we are obviously using sharp null hypotheses about the values of the coefficients of omitted variables.

deviation σ, $0 < \sigma < \infty$ which has an unknown value. The two hypotheses which Jeffreys considers are:

$$H_1 : \lambda = 0 \text{ and } H_2 : \lambda \neq 0. \tag{2.2}$$

As regards prior *pdf*'s, under H_1 Jeffreys utilizes

$$p(\sigma|H_1) \propto 1/\sigma \qquad 0 < \sigma < \infty \tag{2.3}$$

Under H_2, Jeffreys (1967, p. 268) remarks that, "From consideration of similarity it [the prior *pdf* for λ under H_2] must depend on σ, since there, is nothing in the problem except σ to give a scale for λ". His prior under H_2 is

$$p(\lambda,\sigma)d\lambda d\sigma \propto f\left(\frac{\lambda}{\sigma}\right)\frac{d\lambda}{\sigma}\;\frac{d\sigma}{\sigma} \tag{2.4}$$

where $\int_{-\infty}^{\infty} f(\lambda/\sigma)\, d\lambda/\sigma = 1$. Then with prior odds 1:1, the posterior odds ratio, K_{12} is:

$$K_{12} = \frac{\int_0^\infty \sigma^{-n-1} \exp\{-n(\bar{y}^2+\hat{\sigma}^2)/2\sigma^2\}d\sigma}{\int_{-\infty}^{\infty}\int_0^\infty F(\lambda/\sigma)\,\sigma^{-n-2}\exp\{-n[(\lambda-\bar{y})^2+\hat{\sigma}^2]/2\sigma^2\}d\sigma d\lambda} \tag{2.5}$$

where $\bar{y} = \Sigma_{i=1}^n y_i/n$ and $n\hat{\sigma}^2 = \Sigma_{i=1}^n (y_i-\bar{y})^2$.

From detailed consideration of (2.5) in the case $n=1$ in which no decision regarding H_1 and H_2 can be made ($K_{12} = 1$), Jeffreys finds "that the consideration that one observation shall give an indecisive result is satisfied if $f(v)$ [with $v = \lambda/\sigma$] is any even function with integral 1." (p. 269). Further, the condition that $K_{12}=0$ for $n \geq 2$ when $\hat{\sigma}=0$ and $y \neq 0$ requires that the denominator of (2.5) diverge. This will occur if and only if $\int_0^\infty f(v)v^{n-1}dv$ diverges (p.269). As Jeffreys notes, "the simplest function satisfying this condition for $n>1$ and also satisfying (3) [$\int_{-\infty}^{\infty} f(v)dv = 1$] is $f(v) = 1/\pi(1+v^2)$." Thus his form for $f(\lambda/\sigma)$ is

$$f\left(\frac{\lambda}{\sigma}\right)\frac{d\lambda}{\sigma} = \frac{1}{\pi}\;\frac{1}{1+\lambda^2/\sigma^2}\;\frac{d\lambda}{\sigma} \qquad -\infty < \lambda < \infty \tag{2.6}$$

a *pdf* in the univariate Cauchy form centered at zero. With respect to this point, Jeffreys (1967, p.251) states, "We must... say that the mere fact that it has been suggested that λ is zero corresponds to some presumption that it is fairly small". After pointing to unsatisfactory features of a normal prior *pdf* for λ,[3] he writes, "The chief advantage of the form [(2.6)] what we have chosen is that in any significance test it leads to the conclusion that if the null hypothesis [$\lambda = 0$] has a small posterior probability, the posterior probability of the parameters is nearly the same as in the estimation problem. Some difference remains but it is only a trace". (p. 273).

When (2.6) is substituted in (2.5) and the integrations are performed, approximately in terms of the denominator, Jeffreys obtains (1967, p. 272):

$$K_{12} \doteq (\pi\nu/2)^{1/2} / (1 + t^2/\nu)^{(\nu-1)/2} \tag{2.7}$$

where $\nu = n\text{-}1$ and $t = \sqrt{n}\, y/s$, with $s^2 = \Sigma_{i=1}^{n} (y_i\text{-}\bar{y})^2/\nu$ and the error of the approximation "is of the order of $1/n$ of the whole expression". Also Jeffreys (p.274) provides an exact expression for K_{12}. Shown below are values of K_{12} for selected values of ν and t^2 taken from Jeffreys's table (p.439):

From Table 2.1, it is seen that when $\nu = 20$, $K_{12} = 1$ when $t^2 = 4.0$ while for $\nu = 5{,}000$, $K_{12} = 1$ when $t^2 = 9$. It is thus seen that as ν increases in value, a larger value of t^2 is required for indifference ($K_{12} = 1$) between H_1 and H_2. This corresponds to a sampling theorist's usual lowering of the significance level as ν grows in value and also bears a direct relationship to Lindley's Paradox (1957). Also note that in contrast to DeGroot's (1973) result, the tail area or "p-value" associated with the t-value is *not* equal to the posterior probability on the null hypothesis[4]. For example, with $\nu = 20$ and $t = 2.0$, the "p-value" is approximately .025 and yet $K_{12} = 1$ or the posterior probability on $H_1 : \lambda = 0$ is ½. Finally, as Jeffreys (1967, p. 272) remarks, the variation of K_{12} with t is much more important than its variation with ν. For moderately large ν, $K_{12} \doteq (\pi\nu/2)^{1/2}\exp(-t^2/2)$, from which the dependence of K_{12} on ν and t is clearly seen.

In a brief treatment of regression, Jeffreys (1967, pp. 324-326) remarks that "...The whole of the tests related to the normal law of error can be adapted immediately to tests concerning the introduction of a new function to represent a series of measures". (p. 325). He considers the important special case for which the hypothesis is that an added term's coefficient is equal to zero and points out that (2.7) is the approximate posterior odds ratio for this problem where t is the usual t-statistic relating to the added term's coefficient and ν is the degrees of freedom associated with the t-statistic. Below it will be seen that Jeffreys's result is included in our general results as a special case.

[3] Jeffreys (1967, p.273) points out that if the prior *pdf* for $v = \lambda/\sigma$ were $p(v) \propto \exp[-cv^2]$, where c is some given positive constant, the posterior odds ratio for $\lambda = 0$ and $\lambda \neq 0$ "... would never be less than some positive function of n [the sample size] however closely the observations agreed among themselves". Also, on this same page he points out a second defect of this normal form for the prior *pdf*.

[4] It appears that DeGroot (1973) obtains his result that the tail area associated with a sampling theory test statistics's value is equal to the posterior probability on the null hypothesis by use of a very special prior *pdf* on his parameter θ. His prior probabilities on θ's possible values are fixed even though a given departure of θ from its null value of zero implies differing departures of the underlying location parameter's value from zero as n, the sample size changes.

TABLE 2.1

Values of t^2 Associated with Corresponding Values of K_{12} and $\nu = n\text{-}1$ from (2.7)

	K_{12}				
ν	1	$10^{-1/2}$	10^{-1}	$10^{-3/2}$	10^{-2}
9*	3.5	7.7	13.3	...	...
15	3.8	7.1	11.1	15.9	21.5
20	4.0	7.0	10.6	14.5	18.9
50	4.6	7.4	10.0	12.8	16.0
100	5.2	7.7	10.3	12.8	15.5
200	5.7	8.2	10.7	13.1	15.6
500	6.8	9.1	11.4	13.8	16.2
1,000	7.4	9.7	12.0	14.3	16.6
2,000	8.1	10.4	12.7	15.0	17.3
5,000	9.0	11.3	13.6	15.9	18.2
10,000	9.7	12.0	14.3	16.6	18.9
50,000	11.3	13.6	15.9	18.2	20.5
100,000	12.0	14.3	16.6	18.9	21.2

3. POSTERIOR ODDS RATIOS FOR SELECTED REGRESSION HYPOTHESES

Let our regression model for the nx1 observation vector $\mathbf{y}$ be:

$$\mathbf{y} = \alpha\iota + X\beta + \mathbf{u} \tag{3.1}$$

* For $\nu = 9$, Jeffreys has used his exact result for K_{12} to compute the following t^2 values: 3.8 for $K_{12} = 1$, 7.7 for $K_{12} = 10^{-1/2}$, and 13.1 for $K_{12} = 10^{-1}$. It is seen that the exact results are in good agreement with the approximate results even though $\nu = 9$ is small. Jeffreys (1967, p. 439) tabulates exact values for $\nu = 1,2,3,\ldots,9$.

where ι is an nx1 vector with all elements equal to one, α and β are a scalar parameter and a kx1 vector of parameters with unknown values, $(\iota{:}X)$ is an nx$(k+1)$ given matrix of rank $k+1$ and $\mathbf{u}$ is an $n+1$ vector of error terms. It is assumed that the variables in X are measured in terms of deviations from their respective sample means and thus $\iota' X = \mathbf{0}'$ Further, the elements of $\mathbf{u}$ are assumed independently drawn from a normal population with zero mean and finite variance σ^2 with unknown value.

We initially consider the following two hypotheses:

$$H_1 : \beta = 0, -\infty < \alpha < \infty \text{ and } 0 < \sigma < \infty \tag{3.2}$$

$$H_2 : \beta \neq \mathbf{0}, -\infty < \alpha < \infty \text{ and } 0 < \sigma < \infty \tag{3.3}$$

The likelihood functions under these two hypotheses are given by:

$$\begin{aligned} p(\mathbf{y}|\alpha,\sigma,H_1) &\propto \sigma^{-n}\exp\{-(\mathbf{y}-\alpha\iota)'(\mathbf{y}-\alpha\iota)/2\sigma^2\} \\ &\propto \sigma^{-n}\exp\{-[\nu_1 s_1^2 + n(\alpha y)^2]/2\sigma^2\} \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} p(\mathbf{y}|\alpha,\beta,\sigma,H_2) &\propto \sigma^{-n}\exp\{-(\mathbf{y}-\alpha\iota-X\beta)'(\mathbf{y}-\alpha\iota-X\beta)/2\sigma^2\} \\ &\propto \sigma^{-n}\exp\{-[\nu_2 s_2^2 + n(\alpha-\bar{y})^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]/2\sigma^2\} \end{aligned} \tag{3.5}$$

where the proportionality constant is $(2\pi)^{-n/2}$ in each case,

$$y = \Sigma_{i=1}^n y_i/n,\ \nu_1 s_1^2 = \Sigma_{i=1}^n (y_i y)^2,\ \nu_1 = n-1$$
$$\hat{\beta} = (X'X)^{-1}X'\mathbf{y},\ \nu_2 s_2^2 = (\mathbf{y}y\iota-X\hat{\beta})'(\mathbf{y}y\iota-X\hat{\beta}), \text{ and } \nu_2 = n-k-1.$$

The following prior assumptions will be utilized in obtaining a posterior odds ratio. First we place equal prior probabilities of ½ on both hypotheses and thus the prior odds ratio is 1:1. Second, under H_1 we employ a diffuse prior distribution for α and σ, that is,

$$p(\alpha,\sigma|H_1) \propto 1/\sigma \qquad -\infty < \alpha < \infty \text{ and } 0 < \sigma < \infty \tag{3.6}$$

Under H_2 we utilize the following prior *pdf*

$$p(\alpha,\beta,\sigma|H_2) \propto f(\beta|\sigma)/\sigma \qquad -\infty < \alpha < \infty \text{ and } 0 < \sigma < \infty \tag{3.7a}$$

with

$$f(\beta|\sigma) = c|X'X/n\sigma^2|^{1/2}/(1 + \beta'X'X\beta/n\sigma^2)^{(k+1)/2} \quad -\infty < \beta_i < \infty \quad i = 1,2,\ldots k \tag{3.7b}$$

where $c = \Gamma[(k+1)/2]/\pi^{(k+1)/2}$.

In (3.6) and (3.7a) the factors of proportionality are assumed the same. Further, in (3.7b) it has been assumed that the prior *pdf* for β given σ is in the form of a k-dimensional multivariate Cauchy probability density function with zero location vector and matrix $X'X/n$, a matrix suggested by the form of the information matrix.

The posterior odds ratio, K_{12} for H_1 and H_2 with the prior odds ratio 1:1, is:

$$K_{12} = \frac{\int p(\mathbf{y}|\alpha,\sigma,H_1)p(\alpha,\sigma|H_1)d\alpha d\sigma}{\int p(\mathbf{y}|\alpha,\beta,\sigma,H_2)p(\alpha,\beta,\sigma|H_2)d\alpha d\beta d\sigma} \tag{3.8}$$

Explicitly, the integration in the numerator of (3.8) is performed as follows. The integral to be evaluated is:

$$I_N = \int_0^\infty \int_{-\infty}^\infty \sigma^{-(n+1)}\exp\{-[\nu_1 s_1^2 + n(\alpha-\bar{y})^2]/2\sigma^2\}d\alpha d\sigma$$

Using properties of the univariate normal *pdf*, integrate with respect to α to obtain:

$$\begin{aligned} I_N &= (2\pi/n)^{1/2} \int_0^\infty \sigma^{-(\nu_1+1)} \exp\{-\nu_1 s_1^2/2\sigma^2\}d\sigma \\ &= (2\pi/n)^{1/2}\Gamma(\nu_1/2)(2/\nu_1 s_1^2)^{\nu_1/2}/2 \end{aligned} \tag{3.9a}$$

where the integration over σ was performed by utilizing well-known properties of the inverted gamma *pdf*—see, e.g. Zellner (1971, p. 371)—.

The integral in the denominator of (3.8), denoted by I_D will be evaluated as follows:

$$\begin{aligned} I_D = \int & f(\beta|\sigma,H_2)\sigma^{-(n+1)}\exp\{-[\nu_2 s_2^2 + n(\alpha-\bar{y})^2 \\ & + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]/2\sigma^2\}d\alpha d\beta d\sigma \end{aligned}$$

with $f(\beta|\sigma)$ given in (3.7b). First integrate over α using properties of the univariate normal *pdf* to obtain:

$$I_D = (2\pi/n)^{1/2}\int f(\beta|\sigma,H_2)\sigma^{-n}\exp\{-[\nu_2 s_2^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})]/2\sigma^2\}d\beta d\sigma.$$

On inserting $f(\beta|\sigma,H_2)$ from (3.7a) and performing the integration with respect to the elements of β approximately,[5]

$$I_D \doteq (2\pi/n)^{1/2}c(2\pi/\nu_2)^{k/2}\int_0^\infty \frac{1}{\sigma^n}\frac{1}{\{1+\hat{\beta}'X'X\hat{\beta}/n\sigma^2\}^{(k+1)/2}}\exp\{-\nu_2 s_2^2/2\sigma^2\}d\sigma$$

Then,

$$I_D \doteq (2\pi/n)^{1/2}c(2\pi/\nu_2)^{k/2}\Gamma(\nu_1/2)1/2(2/\nu_2 s_2^2)^{\nu 1/2}/(1+\hat{\beta}'X'X\hat{\beta}/\nu_2 s_2^2)^{(k+1)/2} \quad (3.9b)$$

where the integration over σ has beeen performed approximately.

Using the results in (3.9a) and (3.9b), the approximate posterior odds ratio for H_1 vs. H_2 is given by:

$$K_{12} \doteq (1/c)(\nu_2/2\pi)^{k/2}(\nu_2 s_2^2/\nu_1 s_1^2)^{\nu 1/2}(1+\hat{\beta}'X'X\hat{\beta}/\nu_2 s_2^2)^{(k+1)/2}$$

$$= a(\nu_2/2)^{k/2}(\nu_2 s_2^2/\nu_1 s_1^2)^{(\nu 2-1)/2} \quad (3.10)$$

with $a=\pi^{1/2}/\Gamma[(k+1)/2]$, since $\nu_2 s_2^2 + \hat{\beta}'X'X\hat{\beta} = \nu_1 s_1^2$. Alternatively, K_{12} in (3.10) can be expressed as:

$$K_{12} \doteq a(\nu_2/2)^{k/2}/[1+(k/\nu_2)F_{k,\nu_2}]^{(\nu_2-1)/2} \quad (3.11)$$

or

$$K_{12} \doteq a(\nu_2/2)^{k/2}(1-R^2)^{(\nu_2-1)/2} \quad (3.12)$$

where $F_{k,\nu_2} = \hat{\beta}'X'X\hat{\beta}/ks_2^2$ and $R^2 = \hat{\beta}'X'X\hat{\beta}/(\nu_2 s_2^2+\hat{\beta}'X'X\hat{\beta})$, the usual "F-statistics" and the squared sample multiple correlation coefficient, respectively. Further, a large sample approximation to $-2\ell n K_{12}$ is given by:

$$-2\ell n K_{12} \doteq \chi_k^2 - k\ell n \nu_2 \quad (3.13)$$

where $\chi_k^2 = \hat{\beta}'X'\hat{\beta}/s_2^2$.[6]

We now consider a hypothesis relating to a subvector of β in (3.1). Rewrite (3.1) as

$$\mathbf{y} = \alpha\iota + X_1\beta_1 + X_2\beta_2 + \mathbf{u} \quad (3.14)$$

where $X = (X_1\vdots X_2)$ with X_1 and X_2 k_1x1 and k_2x1 vectors, respectively and $k_1+k_2=k$. All other assumptions made in connection with (3.1) apply to (3.14). For convenience, we shall reparametrize (3.14) as follows:

$$\mathbf{y} = \alpha\iota + X_1\eta + V\beta_2 + \mathbf{u} \quad (3.15)$$

where $V = [I-X_1(X_1'X_1)^{-1}X_1']X_2$ and $\eta = \beta_1+(X_1'X_1)^{-1}X_1'X_2\beta_2$. Note that $X_1'V = 0$

A posterior odds ratio relating to the following two hypotheses will be derived:

$$H_A : \beta_2 = \mathbf{0} \quad (3.16)$$

and

$$H_B : \beta_2 \neq \mathbf{0} \quad (3.17)$$

with α, the elements of η and σ unrestricted under both hypotheses.

The likelihood functions under these two hypotheses are given by:

$$p(\mathbf{y}|\alpha,\sigma,\eta,\mathrm{H}_A) \propto \sigma^{-n}\exp\{\nu_A s_A^2 + n(\alpha-\bar{y})^2 + (\eta-\hat{\eta})'X_1'X_1(\eta-\hat{\eta})]/2\sigma^2\} \quad (3.18)$$

and

$$p(\mathbf{y}|\alpha,\sigma,\eta,\beta_2,H_B) \propto \sigma^{-n}\exp\{-[\nu_B s_B^2 + n(\alpha-\bar{y})^2 + \\ + (\eta-\hat{\eta})'X_1'X_1(\eta-\hat{\eta}) + (\beta_2-\hat{\beta}_2)'V'V(\beta_2-\hat{\beta}_2)]/2\sigma^2\} \quad (3.19)$$

where

[5] This approximate integration can be viewed as finding the mean of $f(\beta|\sigma, H_2)$ a bounded function of β. Cramer (1946, p. 353 ff.) indicates that the error of the approximation is $0(n^{-1})$ in line with Jeffreys's remark cited in Section II. Thus if the posterior odds ratio $K_{12}=I_N/I_D$ and if the integral I_N is evaluated exactly and $I_D=\hat{I}_D+0(n^{-1})$, where $\hat{I}_D$ is the approximate value of I_D, $K_{12}=I_N/[\hat{I}_D+0(n^{-1})]$ or $I_N/\hat{I}_D = K_{12}[1+0(n^{-1})]$ and thus the error in using I_N/I_D is $K_{12}.0(n^{-1})$, as pointed out above bv Jeffreys.

[6] To obtain (3.13), write (3.11) as $K_{12} \doteq a(\nu_2/2)^{k/2}\exp\{\nu^2-1)/2\ \ell n\ [1+(k/\nu^2)F_{k,\nu_2}]\}$ and expand the logarithmic factor in the exponential as $\ell n(1+x) \doteq x$. The result is $K_{12}\doteq a(\nu_2/2)^{k/2}\exp\{-kF_{k,\nu_2}/2\}$. Then $-2\ell n K_{12} \doteq \chi_k^2 - k\ell n \nu_2$, where $\chi_k^2 = kF_{k,\nu_2}$ and terms not depending on ν_2 have been dropped in this large-ν_2 approximation. Further, under $\beta=0$ the approximate sampling *pdf* for $-2\ell n K_{12}$ can be obtained from that of χ_k^2. Also, again under $\beta = \mathbf{0}$ the approximate cumulative sampling *pdf* for K_{12} in (3.11) can be obtained from that of F_{k,ν_2}. That is, since K_{12} is a one-to-one monotonic function of F_{k,ν_2} for fixed k and ν_2, $\Pr(F_{k,\nu_2} > x)$ $= \Pr(K_{12} < x')$, where x' is the value of K_{12} associated with $F_{k,\nu_2} = x$.

$$\bar{y} = \Sigma_{i=1}^{n} y_i/n,\ \hat{\eta} = (X_1'X_1)^{-1}X_1'\mathbf{y}, \qquad \nu_A s_A^2 = (\mathbf{y}-\bar{y}\iota - X_1\hat{\eta})'(\mathbf{y}-\bar{y}\iota - X_1\hat{\eta}),$$

$$\nu_A = n-k_1-1,\ \hat{\beta}_2 = (V'V)^{-1}V'\mathbf{y}, \qquad \nu_B s_B^2 = (\mathbf{y}-\bar{y}\iota - X_1\hat{\eta} - V\hat{\beta}_2)'(\mathbf{y}-\bar{y}\iota - X_1\hat{\eta} - V\hat{\beta}_2)$$

and $\nu_B = n-k_1-k_2-1$.

Under H_A, we employ the following diffuse prior *pdf* for the parameters:

$$p_A(\alpha,\sigma,\eta\,|\,H_A) \propto |X_1'X_1|^{1/2}/\sigma \quad -\infty < \alpha,\eta_i < \infty \quad i = 1,2,\ldots,k_1 \tag{3.20}$$

$$0 < \sigma < \infty$$

while under H_B the prior *pdf* is:

$$p_B(\alpha,\sigma,\eta,\beta_2\,|\,H_B) \propto |X_1'X_1|^{1/2}h(\beta_2\,|\,\sigma)/\sigma \tag{3.21a}$$

$$-\infty < \alpha,\ \eta_i < \infty \qquad i = 1,2,\ldots,k_1$$
$$0 < \sigma < \infty$$

with

$$h(\beta_2\,|\,\sigma) = c_B|V'V/n\sigma^2|^{1/2}/(1+\beta_2'V'V\beta_2/n\sigma^2)^{(k_2+1)/2} \tag{3.21b}$$

$$-\infty < \beta_{2i} < \infty \qquad i = 1,2,\ldots,k_2.$$

In (3.20) and (3.21a) the factor of proportionality is taken to be the same. In (3.21b), the prior *pdf* for β_2 given σ is the form of a k_2-dimensional multivariate Cauchy *pdf* with zero location vector and matrix $V'V/n$, a matrix suggested by the form of the information matrix.

The posterior odds ratio, K_{AB} for H_A and H_B with the prior odds ratio 1:1 is:

$$K_{AB} = \frac{\int p(\mathbf{y}\,|\,\alpha,\sigma,\eta,H_A)p_A(\alpha,\sigma,\eta\,|\,H_A)d\alpha d\sigma d\eta}{\int p(\mathbf{y}\,|\,\alpha,\sigma,\eta,\beta_2,H_B)p_B(\alpha,\sigma,\eta,\beta_2\,|\,H_B)d\alpha d\sigma d\eta d\beta_2} \tag{3.22}$$

On applying integration techniques similar to those employed above (see Appendix), the following approximate expression for K_{AB} is obtained:

$$\begin{aligned} K_{AB} &\doteq b(\nu_B/2)^{k_2/2}(\nu_B s_B^2/\nu_A s_A^2)^{(\nu_B-1)/2} \\ &= b(\nu_B/2)^{k_2/2}[(1-R_B^2)/(1-R_A^2)]^{(\nu_B-1)/2} \\ &= b(\nu_B/2)^{k_2/2}/[1+(k_2/\nu_B)F_{k_2\nu_B}]^{(\nu_B-1)/2} \end{aligned} \tag{3.23}$$

where $b = \pi^{1/2}/\Gamma[(k_2+1)/2]$, R_A^2 and R_B^2 are the squared sample multiple correlation coefficientes under H_A and H_B and $F_{k_2\nu_B} = \hat{\beta}_2'V'V\hat{\beta}_2/k_2 s_B^2$, the usual "*F*-statistic". Also, if ν_B is large, the following approximate result is available:

$$-2\ell n K_{AB} \doteq \chi_{k_2}^2 - k_2\ell n\nu_B = \nu_B(R_B^2-R_A^2)/(1-R_B^2) - k_2\ell n\nu_B \tag{3.24}$$

with $\chi_{k_2}^2 = \hat{\beta}_2'V'V\hat{\beta}_2/s_B^2$.

We now consider the following four hypotheses relating to β_1 and β_2 in (3.14), each assumed to have the same prior probability:

$$H_1: \beta_1 = \mathbf{0} \text{ and } \beta_2 = \mathbf{0}, \tag{3.25a}$$

$$H_2: \beta_1 \neq \mathbf{0} \text{ and } \beta_2 \neq \mathbf{0}, \tag{3.25b}$$

$$H_3: \beta_1 \neq \mathbf{0} \text{ and } \beta_2 = \mathbf{0}, \tag{3.25c}$$

and

$$H_4: \beta_1 = \mathbf{0} \text{ and } \beta_2 \neq \mathbf{0}, \tag{3.25d}$$

The posterior odds ratio for H_1 and H_2, K_{12}, given in (3.11) is:

$$K_{12} \doteq a(\nu_2/2)^{k/2}/[1+(k/\nu_2)F_{k\nu_2}]^{(\nu_2-1)/2} \tag{3.26}$$

where $a = \pi^{1/2}/\Gamma[(k+1)/2]$ and $\nu_2 = n-k-1$. This odds ratio has been derived employing the prior assumptions in (3.6) and (3.7), the latter involving a multivariate Cauchy prior *pdf* for β_1 and β_2 given σ. The posterior odds ratio for H_3 and H_2, K_{32} is identical to K_{AB} in (3.23), namely

$$K_{32} = b(\nu_2/2)^{k_2/2}/[1+(k^2/\nu_2)F_{k_2\nu_2}]^{(\nu_2-1)/2} \tag{3.27}$$

where $b = \pi^{1/2}/\Gamma[(k_2+1)/2]$ and $\nu_2 = \nu_B = n-k-1$. K_{32} also can be obtained by using the conditional prior *pdf* for β_1 given $\beta_2 = \mathbf{0}$ and σ associated with the multivariate Cauchy *pdf* in (3.7b) under H_2 along with uniform independent priors for α and log σ. Similarly, the posterior odds ratio for H_4 and H_2,K_{42} can be obtained and is:

$$K_{42} \doteq q(\nu_2/2)^{k_1/2}/[1 + (k_1/\nu_2)F_{k_1\nu_2}]^{(\nu_2-1)/2} \tag{3.28}$$

where $q = \pi^{1/2}/\Gamma[(k_1+1)/2]$. Last, from (3.27) and (3.28), the posterior odds ratio for H_3 and H_4,K_{34} is:

$$K_{34}=K_{32}/K_{42} \doteq g(\nu_2/2)^{(k_2-k_1)/2} \left(\frac{1+(k_1/\nu_2)F_{k_1\nu_2}}{1+(k_2/\nu_2)F_{k_2\nu_2}}\right)^{(\nu_2-1)/2} \tag{3.29}$$

where $g = \Gamma[(k_1+1)/2]/\Gamma[(k_2+1)/2]$.

The posterior odds ratios in (3.26)-(3.29) can be helpful in screening sets of variables, X_1 and X_2 for inclusion a in regression in situations in which there is little prior information and the initial presumption is that neither set of variables probably belongs in the regression. A special case of the above analysis is one in which X_1 and X_2 are vectors and thus β_1 and β_2 are scalars. In this case, we are screening individual variables por possible inclussion in the regression. Further, elaboration of the hypotheses in (3.25) to relate individual coefficients is possible and would lead to posterior odds ratios useful in determining which individual variables to include in a regression.

To gain greater familiarity with the odds ratios derived above, we now turn to consider some numerical evaluations of them.

4. NUMERICAL EVALUATION OF SELECTED ODDS RATIOS

In this Section, we provide some numerical evaluations of the odds ratios derived in Section III. First, note that when $k = 1$, the posterior odds ratio K_{12} in (3.11) for the hypotheses $\beta = \mathbf{0}$ and $\beta \neq \mathbf{0}$ reduces to $K_{12} \doteq (\pi\nu_2/2^{1/2}/(1 + t^2/\nu^2)^{(\nu_2-1)/2}$, with $\nu^2 = n - 2$ which is exactly in the form of Jeffreys's odds ratio in (2.7). Thus the numerical results in Table 2.1 apply directly to the case of simple regression. From Table 2.1, it is seen that for $\nu_2 = 20$, $K_{12} \doteq 1$ when $t^2 = 4.0$ where $t^2 = \hat{\beta}^2\Sigma(x_i - \bar{x})^2/s^2$ is the square of the usual t-statistic. Since $r^2 = t^2/(\nu^2 + t^2)$, a value of $r^2 = 1/6$ corresponds to $t^2 = 4.0$ and $K_{12} \doteq 1$ for $\nu_2 = 20$. For $\nu_2 = 5{,}000$ and $t^2 = 9.0$ (or $r^2 = .0018$), $K_{12} \doteq 1$. Thus indifference, ($K_{12} \doteq 1$) is achieved for a larger value of t^2 (or a lower value of r^2) with $\nu_2 = 5{,}000$ as compared with $\nu_2 = 20$. For $\nu_2 = 20$, $K_{12} \doteq 1/100$, that is the odds are 100:1 against $\beta = 0$ when $t^2 = 18.9$ or $r^2 = .486$. For $\nu_2 = 5{,}000$, $K_{12} \doteq 1/100$ when $t^2 = 18.2$ or $r^2 = .00377$. Thus with $\nu_2 = 5{,}000$, a value of $t^2 = 18.2$ (or equivalently, $r^2 = .00377$) strongly favors the hypotheses $\beta \neq 0$. Since values of ν_2 in the vicinity of several thousand are frequently encountered in analyses of cross-section or survey data, these results are relevant for applied work. In particular, they point (a) the need for absolutely larger t-values for indifference ($K_{12} = 1$) as ν_2 increases and (b) recognition that for large values of ν_2, small values of r^2 can be consistent with strong evidence *against* $\beta = 0$. These results, it must be emphasized, apply in situations in which we have little prior information about β's value under the hypotheses $\beta \neq 0$. If more information is available, suitable prior *pdf*'s reflecting it would have to be introduced, as pointed out by Jeffreys (1967, p. 252).

TABLE 4.1
Values of R^2 and F_{k,ν_2} Associated with
Particular Values of K_{12} and k in
(3.12) for $\nu_2 = 20$ and $\nu_2 = 100$*
A. $\nu_2 = 20$

k	Value of:	K_{12} 1	 $10^{-1/2}$	 10^{-1}	 $10^{-3/2}$	 10^{-2}	.01 and .05 Critical Values of F and Associated R^2's .01	 .05
1	R^2	.16	.26	.35	.42	.49	.29	.18
	$F_{1,20}$	4.0	7.0	10.6	14.5	18.9	8.10	4.35
2	R^2	.27	.35	.43	.49	.55	.37	.26
	$F_{2,20}$	3.7	5.5	7.5	9.7	12.3	5.85	3.49
3	R^2	.35	.42	.48	.54	.60	.43	.32
	$F_{3,20}$	3.5	4.8	6.3	8.0	9.9	4.94	3.10
4	R^2	.40	.47	.53	.58	.63	.47	.36
	$F_{4,20}$	3.4	4.4	5.7	7.0	8.6	4.43	2.87
5	R^2	.45	.51	.57	.61	.66	.51	.40
	$F_{5,20}$	3.2	4.2	5.2	6.4	7.7	4.10	2.71
6	R^2	.48	.54	.59	.64	.68	.54	.44
	$F_{6,20}$	3.1	3.9	4.9	5.9	7.1	3.87	2.60
				B. $\nu_2 = 100$				
1	R^2	.050	.072	.093	.11	.13	.065	.038
	$F_{1,100}$	5.2	7.7	10.3	12.8	15.5	6.90	3.94
2	R^2	.089	.11	.13	.15	.17	.088	.058
	$F_{2,100}$	4.9	6.2	7.5	8.8	10.3	4.82	3.09
3	R^2	.12	.14	.16	.18	.20	.11	.075
	$F_{3,100}$	4.6	5.5	6.4	7.4	8.3	3.98	2.70
4	R^2	.15	.17	.19	.21	.23	.12	.090
	$F_{4,100}$	4.4	5.1	5.9	6.6	7.3	3.51	2.46
5	R^2	.18	.20	.21	.23	.25	.14	.10
	$F_{5,100}$	4.3	4.9	5.5	6.1	6.7	3.20	2.30
6	R^2	.20	.22	.24	.25	.27	.15	.12
	$F_{6,100}$	4.2	4.7	5.2	5.7	6.2	2.99	2.19

*Note that $F_{k,\nu_2} = (\nu_2/k)R^2/(1-R^2)$, with $\nu_2 = n$-k-1.

In Table 4.1, we have evaluated the posterior odds ratio K_{12} for $H_1 : \beta = 0$ vs. $H_2 : \beta \neq 0$ for $\nu_2 = 20$ and $\nu_2 = 100$, where $\nu_2 = n\text{-}k\text{-}1$ for selected values of k, the number of elements in β and selected values of R^2, the sample squared multiple correlation coefficient. Also shown in the table are values of associated F-statistics, $F_{k,\nu_2} = (\nu_2/k)R^2/(1\text{-}R^2)$, and .01 and .05 critical values of the F statistic as well as the R^2 values associated with these critical values. From the first line of Table 4.1, we see that for $k = 1$ and $\nu_2 = n\text{-}k\text{-}1 = 20$, $K_{12} \doteq 1$ when $R^2 = .16$ and $F_{1,20} = t^2_{20} = 4.00$. Note that for these conditions the sampling theorists's .05 critical value of F is $F_{1,20}(.05) = (2.086)^2 = 4.35$ with an associated $R^2 = .18$. Thus the 5% F value is somewhat larger than the Bayesian indifference ($K_{12} = 1$) value of 4.0. Alternatively, an $R^2 = .16$ leads to $K_{12} \doteq 1$ while an $R^2 = .18$ is associated with the sampling theorists's .05 critical value of F. On the other hand, a .01 critical value of F is 8.10, with an associated $R^2 = .29$ which is far from the F value 4.0, or $R^2 = .16$ which yields $K_{12} \doteq 1$.

TABLE 4.2

Values of χ^2_k and Associated R^2's for $K_{12} \doteq 1$ Using Approximate Formula (3.13) for $\nu_2 = 20$ and $\nu_2 = 100$ and Selected Values for k

	$\nu_2 = 20$		$\nu_2 = 100$		.05 and .01 Critical Values of χ^2_k	
k	χ^2_k	R^{2*}	χ^2_k	R^{2*}	$\chi^2_k(.05)$	$\chi^2_k(.01)$
1	3.00	.13	4.6	.044	3.84	6.63
2	5.99	.23	9.2	.084	5.99	9.21
3	8.99	.31	13.8	.12	7.81	11.30
4	12.0	.37	18.4	.16	9.49	13.30
5	15.0	.43	23.0	.19	11.10	15.10
6	18.0	.47	27.6	.22	12.60	16.80

* Note that $\lambda^2_k = \hat{\beta}'X'X\hat{\beta}/s^2$ and $R^2 = \chi^2_k/(\nu_2 + \chi^2_k)$, where $\nu_2 = n\text{-}k\text{-}1$.

In Table 4.2 values of χ^2_k and associated values of R^2 which correspond to $K_{12} = 1$ have been tabulated for $\nu_2 = 20$ and $\nu_2 = 100$ and $k = 1,2,\ldots,6$, to gain some information on the quality of the approximation in (3.13). The entries in Table 4.2 have been computed from the large sample approximate formula (3.13), that is $-2\ell n K_{12} \doteq \chi^2_k - k\ell n \nu_2$. Also shown in Table 4.2 are the .05 and .01 critical values of χ^2_k. For $k = 1$ and $\nu_2 = 20$, the indifference values of χ^2_1 and R^2 are 3.00 and .13, respectively. The latter value can be compared with the more accurate indifference value of $R^2 = .16$ given in Table 4.1. The difference in these values arises because the results in Table 4.2 are based on a cruder approximation than those in Table 4.1. For $\nu_2 = 100$, the corresponding indifference ($K_{12} = 1$) values of R^2 in Tables 4.1 and 4.2 are fairly similar in value. Also, from Table 4.2, the relation of the crude indifference values of χ^2_k can be compared with the .05 and .01 sampling theory critical values of χ^2_k. For $\nu_2 = 100$, it is seen that for $k = 1$, the indifference value of χ^2_1, namely 4.6 falls between the .05 critical value, 3.84 and the .01 critical value, 6.63. For $k > 2$ and $\nu_2 = 100$, the Bayesian indifference values of χ^2_k are all larger than the .01 critical values of χ^2_k.

As regards other posterior odds ratios derived in Section III, it is the case that the numerical results in Tables 2.1, 4.1 and 4.2, can be utilized to evaluate them provided that the degrees of freedom and k parameters are suitably reinterpreted. For example, the expressions for K_{AB} in (3.23) and (3.24) can be evaluated if in using the tables, k is replaced by k_2, $F_{k,\nu}{}^2$ is replaced by F_{k_2,ν_2} (note $\nu_2 = \nu_B = n\text{-}k\text{-}1$) and χ^2_k is replaced by χ^2_{k2}. Similarly, the odds ratios K_{32} and K_{42} in (3.27) and (3.28), respectively can be implemented using the results in the tables by similar redefinitions. Last, the odds ratio, K_{32} in (3.29) can be evaluated from results given in Table 4.1 by use of $K_{32} = K_{32}/K_{42}$, where values of K_{32} and K_{42} can be obtained from entries in Table 4.1. Finally, it is to be noted that in the expression for K_{32} in (3.29), there is an interesting allowance for the possibly differing numbers of parameters in β_1 and β_2.

5. SUMMARY AND CONCLUDING REMARKS

In this paper we have derived approximate posterior odds ratios for sharp null hypotheses which are frequently encountered in regression analyses. These posterior odds ratios are appropriate when little is known regarding parameter values and special attention is given to specific values, e.g. zero values of the regression coefficients. With slight modifications, other special values can be incorporated in the analysis by reparametrizing to convert the null hypotheses to involve zero values. In our work we have employed asymptotic expansions of certain integrals which are very convenient, yield results which can be compared directly with sampling theory analyses, and are quite accurate, as shown in Jeffreys's work. With some extra computational cost, a numerical integration approach, suggested by Dickey (1971) could be applied to obtain slightly more accurate results.

In line with Jeffreys's, Lindley's and some others's previous results, we have found that sampling theorists's usual .05 critical values of test statistics can be far from a Bayesian posterior odds indifference value of one under a variety of circumstances. Whether this finding is interpreted as a systematic flaw in sampling theory practice is of course critically dependent on the nature of the usually implicit loss structure used in sampling theory testing. Cases in which sampling theorists mechanically employ a 5% significance level no matter what the sample size and/or the number of parameters are interpreted as flawed analyses. If sampling theorists and Bayesians carefully consider the underlying loss structure in choosing between or among hypotheses, the above analysis indicates that there can be a compatibility between Bayesian and sampling theory results in testing but, of course their interpretations will differ radically.

While, as pointed out above there can be some degree of compatibility between Bayesian and sampling theory testing results, the direct interpretation of sample evidence, as reflected in F statistics or R^2 values in terms of posterior odds ratios stands in marked contrast to sampling theorists's and others's unclear interpretations of sample evidence in terms of "p-values", and/or values of R^2 or of $\bar{R}^2$, the adjusted coefficient of determination. As mentioned above, a p-value associated with the value of a test statistic is not at all an accurate measure of the posterior probability associated with a null hypothesis. However, it should be noted that most of the posterior odds ratios derived above are monotonically increasing functions of the p-values associated with t or F statistics involved in the posterior odds ratios. Thus there is some rationale for considering p-values; however, since posterior odds ratios have a direct interpretation and explicitly reflect the prior information employed, their use is preferable to the use of p-values. Also, posterior odds ratios on alternative hypotheses can be employed, as described below to average estimates (and predictions) over alternative hypotheses when posterior odds ratios do not yield a clear-cut choice of a particular hypothesis.

In terms of the hypotheses considered above, it is possible to use their associated posterior odds ratios to obtain optimal (relative to quadratic loss) Bayesian "pre-test" point estimates --see Zellner and Vandaele (1974, pp. 640-641). For example with respect to the hypotheses $H_1 : \beta = \mathbf{0}$ and $H_2 : \beta \neq \mathbf{0}$, the point estimate that is optimal relative to quadratic loss is $\bar{\beta} = P_1\mathbf{0} + (1-P_1)\tilde{\beta} = (1-P_1)\tilde{\beta} = (1+K_{12})^{-1}\tilde{\beta}$, where P_1 is the posterior probability for H_1, $K_{12} = P_1/(1-P_1)$ is the posterior odds ratio for H_1 and H_2, and $\tilde{\beta}$ is the posterior mean for β under H_2. With the prior pdf (3.7) which we have employed under H_2, $\tilde{\beta}$ will be close to $\tilde{\beta}$, the least-squares estimate. Thus $\bar{\beta} \doteq (1+K_{12})^{-1}\hat{\beta}$ where K_{12} given in (3.11) is a function of the usual F statistic. Also note that the "shrinkage factor", $(1+K_{12})^{-1}$ is between zero and one with a value near zero when K_{12} is large and a value near one when K_{12} is small. This shrinkage factor can be compared with others which have appeared in the sampling theory literature —see Zellner and Vandaele (1975, p. 639).

Finally, it would be interesting to compare the posterior odds ratios derived above with others based on more informative prior distributions.

APPENDIX

Herein we evaluate the integrals appearing in equation (3.22) of the text. The integral in the numerator, denoted by I_N is:

$$I_N \propto \int |X_1'X_1|^{1/2}\sigma^{-(n+1)}\exp\{-[\nu_A s_A^2 + n(\alpha\bar{y})^2 + (\eta-\hat{\eta})'X_1'X_1(\eta-\hat{\eta})]/2\sigma^2\}d\alpha d\eta d\sigma. \tag{A.1}$$

where ν_A, s_A^2, $\bar{y}$ and $\hat{\eta}$ have been defined in the text in connection with (3.18). We can integrate over α and the k_1 elements of η using properties of univariate and multivariate normal pdf's, respectively to obtain:

$$I_N \propto (2\pi)^{(k_1+1)/2}n^{-1/2}\int_0^\infty \sigma^{-\nu_A}\exp\{-\nu_A s_A^2/2\sigma^2\}d\sigma. \tag{A.2}$$

Using properties of the inverted gamma pdf, the integral in (A.2) can be evaluated to yield:

$$I_N \propto (2\pi)^{(k_1+1)/2}n^{-1/2}\Gamma(\nu_A/2)(1/2)(2/\nu_A s_A^2)^{\nu_A/2}. \tag{A.3}$$

The integral in the denominator of (3.22) in the text, denoted by I_D is:

$$I_D \propto \int h(\beta_2|\sigma)|X_1'X_1|^{1/2}\sigma^{-(n+1)}\exp\{-[\nu_B s_B^2 + n(\alpha-\bar{y})^2 + (\eta-\hat{\eta})'X_1'X_1(\eta-\hat{\eta}) + (\beta_2-\hat{\beta}_2)'V'V(\beta_2-\hat{\beta}_2)]/2\sigma^2\}d\alpha d\eta d\sigma \tag{A.4}$$

where $h(\beta_2|\sigma)$ is given in (3.21b) of the text and ν_B, s_B^2, $\hat{\eta}$, $\hat{\beta}_2$ and V have been defined in connection with (3.19). The integration over α and η can be performed exactly using properties of normal distributions to yield:

$$I_D \propto (2\pi)^{(k_1+1)/2}n^{-1/2}\int h(\beta_2|\sigma)\sigma^{-(n-k_1)}\exp\{-[\nu_B s_B^2 + (\beta_2-\hat{\beta}_2)'V'V(\beta_2-\hat{\beta}_2)]/2\sigma^2\}d\beta_2 d\sigma \tag{A.5}$$

The integration over β_2 can be done approximately by noting that is it equivalent to obtaining the expectation of the bounded function $h\,(\beta_2|\sigma)$. Following Jeffreys's approach and also Cramér's (1946, p. 353) approximation results, we have on integrating approximately with respect to the k_2 elements of β_2

$$I_D \propto (2\pi)^{(k+1)/2} n^{-(k_2+1)/2}\, c_B \int_0^\infty (1 + \hat{\beta}_2' V' V \hat{\beta}_2 / n\sigma^2)^{-(k_2+1)/2}$$

$$\cdot\ \sigma^{-(n-k_1)} \exp\{-\nu_B s_B^2/2\sigma^2\} d\sigma$$

Large values of the second two factors in the integrand of this last expression are near $\nu_B s_B^2/n$. If, as Jeffreys (1967, p. 272) does, we substitute $\sigma^2 = \nu_B s_B^2/n$ in the first, slowly varying factor of the integrand, and then integrate with respect to σ, the result is:

$$I_D \propto c_B 2\pi^{(k+1)/2} n^{-(k_2+1)/2} (1 + \hat{\beta}_2' V' V \hat{\beta}_2 / \nu_B s_B^2)^{-(k_2+1)/2} \qquad (A.6)$$

$$\text{x}\, \Gamma\, (\nu_A/2)\, (1/2)\, (2/\nu_B s_B^2)^{\nu_A/2}$$

Then, using (A.3) and (A.6)

$$K_{AB} \doteq \frac{I_N}{I_D} = \frac{1}{c_B} \left(\frac{n}{2\pi}\right)^{k_2/2} \left(\frac{\nu_B s_B^2}{\nu_A s_A^2}\right)^{\nu_A/2} (1 + \hat{\beta}_2' V' V \hat{\beta}_2 / \nu_B s_B^2)^{(k_2+1)/2} \qquad (A.7)$$

Now c_B, the normalizing constant in (3.21b) is

$c_B = \Gamma\,[(k_2+1)/2]/\pi^{(k_2+1)/2}$ and $\nu_A s_A^2 = \nu_B s_B^2 + \hat{\beta}_2' V' V \hat{\beta}_2$. Thus,

$$K_{AB} \doteq b\,(\nu_B/2)^{k_2/2}\, (\nu_B s_B^2/\nu_A s_A^2)^{(\nu_B-1)/2} \qquad (A.8)$$

with $b = \pi^{1/2}/\Gamma[(k_2+1)/2]$ and where in going from (A.7) to (A.8) $(n/2)^{k_2/2}$ has been replaced by $(\nu_B/2)^{k_2/2}$ which to the order of the approximation is equivalent. (A.8) is exactly the expression in (3.23) in the text. Further, on substituting $\nu_B s_B^2/\nu_A s_A^2 = (1-R_B^2)/(1-R_A^2)$ in (A.8), the second line of (3.23) is obtained. Finally, from $\nu_A s_A^2 = \nu_B s_B^2 + \hat{\beta}_2' V' V \hat{\beta}_2$, $\nu_A s_A^2/\nu_B s_B^2 = 1 + \hat{\beta}_2' V' V \hat{\beta}_2/\nu_B s_B^2 = 1 + (k_2/\nu_B) F_{k_2 \nu_B}$, which when utilized in (A.8) gives the third line of (3.23).

ACKNOWLEDGEMENTS

Research financed by income from the H.G.B. Alexander Endowment Fund, Graduate School of Business, U. of Chicago and by National Science Foundation Grant, SOC 77-15662.

REFERENCES

CRAMER, H. (1946) *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

DEGROOT, M.H. (1973) Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio, *J. Amer. Statist. Assoc.* **68**, 966-969.

DICKEY, J.M. (1971) The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters *Ann. Math. Statist.* **42**, 204-223.

— (1975) Bayesian Alternatives to the F-test and Least-Squares Estimates in the Normal Linear Model. In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, (S.E. Fienberg and A. Zellner eds.), 515-554 Amsterdam: North-Holland.

— (1977) Is the Tail Area Useful as an Approximate Bayes factor? *J. Amer. Statist. Assoc.* **72**, 138-142.

GAVER, K.M. and M.S. GEISEL (1974) Discriminating Among Alternative Models: Bayesian and Non-Bayesian Methods, in *Frontiers of Econometrics*, (P. Zarembka ed.) New York: Academic Press.

GEISEL, M.S. (1970) Comparing and Choosing Among Parametric Statistical Models: A Bayesian Analysis with Macroeconomic Application. Ph. D. Thesis. University of Chicago.

JAYNES, E.T. (1976) Confidence Intervals Vs. Bayesian Intervals, in *Foundations of Probability. Theory, Statistical Inference, and Statistical Theories of Science*, (W.L. Harper and C.A. Hooker eds.) 175-213. Dordrecht-Holland: D. Reidel.

JEFFREYS, H. (1957), *Scientific Inference* (2nd ed.) Cambridge: University Press.

— (1967) *Theory of Probability* (3rd rev. ed.), Oxford: University Press.

(1980) "Some General Points in Probability Theory", in *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner ed.) 451-453. Amsterdam: North-Holland.

LEAMER, E.E. (1978) *Specification Searches*, New York: Wiley

LEMPERS, F.B. (1971), *Posterior Probabilities of Alternative Linear Models*, Rotterdam: University Press.

LINDLEY, D.V. (1957) A Statistical Paradox, *Biometrika* **44**, 187-192.

— (1961) The Use of Prior Probability Distributions in Statistical Inference and Decisions, in *Proc. Fourth Berkeley Symp.* (J. Neyman, ed.) 453-468. Berkeley: University of California Press.

SCHWARZ, G. (1978) Estimating the Dimension of a Model, *Ann. Statist.* **6**, 461-464.

THORNBER, E.H. (1966) Applications of Decision Theory to Econometrics, Ph. D. thesis, University of Chicago.

ZELLNER, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley

— and W. VANDAELE (1975) Bayes-Stein Estimators for k-means, Regression and Simultaneous Equation Models, in *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage* (S.E. Fienberg and A. Zellner eds.) 627-653 Amsterdam: North-Holland.

A Bayesian analysis of classical hypothesis testing

JOSÉ M. BERNARDO
Universidad de Valencia

SUMMARY

The procedure of maximizing the missing information is applied to derive reference posterior probabilities for null hypotheses. The results shed further light on Lindley's paradox and suggest that a Bayesian interpretation of classical hypothesis testing is possible by providing a one-to-one approximate relationship between significance levels and posterior probabilities.

Keywords: BAYESIAN INFERENCE; HYPOTHESIS TESTING; INFORMATION MEASURES; LINDLEY'S PARADOX; NON-INFORMATIVE PRIORS; REFERENCE DISTRIBUTIONS.

1. THE PROBLEM

The Bayesian approach to the classical problem of hypothesis testing seems to be clear in principle. Indeed, if one does not have a specific decision problem in mind, to test a null hypothesis H_0 given some data D, i.e. to check whether the data D are compatible with H_0, may well be done by quoting the corresponding posterior probability $p(H_0|D)$ and checking whether or not this is very small.

To produce a posterior probability one needs a prior. If one is interested in an answer which only depends on the data and the model, so that a comparison with classical results is possible, one is bound to use some sort of ''non-informative' or *reference* prior. When both the null hypothesis H_0 and its alternative H_1 have the same dimension, e.g. when both are simple or both composite, a solution is easily obtained. Thus, if both hypotheses, are simple, the widely accepted reference prior is $\pi(H_0) = \pi(H_1) = 1/2$ and the corresponding *reference* posterior probability of the null hypothesis is simply

$$\pi(H_0|D) = \frac{p(D|H_0)}{p(D|H_0)+p(D|H_1)} \tag{1}$$

This seems to behave properly. If both hypotheses are composite, so that the distribution of the data $p(D|\theta)$ may be indexed by some unknown $\theta\epsilon\Theta$ and H_0 is a proper subset of Θ with non-zero measure, one may use a reference prior $\pi(\theta)$ for θ to obtain

$$\pi(H_0|D) = \int_{Ho} \pi(\theta|D)d\theta = \frac{\int_{Ho} p(D|\theta)\pi(\theta)d\theta}{\int_{\Theta} p(D|\theta)\pi(\theta)d\theta} \qquad (2)$$

Again, for a sensible choice of $\pi(\theta)$, this seems to behave properly. Both (1) (2) were proposed by Jeffreys (1939/67), who also sugested

$$\pi(\theta) = i(\theta)^{1/2}$$

where $i(\theta) = -\int p(D|\theta)[\partial^2\{\log p(D|\theta)\}/\partial\theta^2]dD$, as the appropriate choice for $\pi(\theta)$. Maximizing the missing information (see Bernardo, 1979b) provides a general method to derive reference distributions which reduce to (1) and (2) in these cases, and produces Jeffrey's prior under regularity conditions.

It is easy to show, however, that the posterior probability of the null hypothesis $p(H_0|D)$ may be very misleading when H_0 is simple and H_1 composite unless one is very careful with the prior specification. To see this, let $D = \{x_1, x_2,\ldots,x_n\}$ be a random sample from a normal distribution of mean μ and known variance σ^2, let the prior probability that $\mu=\mu_0$, the value of the null hypothesis, be $p\neq 0$; this is necessary to obtain a non-zero posterior probability for H_0. Suppose that the remainder of the prior probability is normally distributed with mean μ_1 and variance σ_1^2 so that

$$p\{H_0\} = p \qquad p\{H_1\} = 1\text{-}p$$
$$p(\mu|H_1) = N(\mu|\mu_1,\sigma_1{}^2)$$

the arithmetic mean of the observation $\bar{x}$ is obviously sufficient and we that

$$p\{H_0|D\} = p\,\{H_0|\bar{x}\} = \frac{p(\bar{x}|H_0)p}{p(\bar{x}|H_0)\,p+p\,(\bar{x}|H_1)\,(1\text{-}p)}$$

$$= \left(1 + \frac{1\text{-}p}{p}\,\frac{p(\bar{x}|H_1)}{p(\bar{x}|H_0)}\right)^{-1} \qquad (3)$$

where

$$p(\bar{x}|H_0) = N(\bar{x}|\mu_0, \sigma^2/n) \;,$$

$$p(\bar{x}|H_1) = \int N(\bar{x}|\mu, \sigma^2/n)\; N(\mu|\mu_1, \sigma_1{}^2)d\mu = N(\bar{x}|\mu_1, \sigma_1{}^2+\sigma^2/n) \;,$$

so that (3) becomes

$$p\{H_0|D\} = \left[\,1 + \frac{1\text{-}p}{p}\left(\frac{\sigma^2/n}{(\sigma_1^2+\sigma_2/n)}\right)^{1/2} \frac{\exp\{(\bar{x}\text{-}\mu_0)^2/(2\sigma^2/n)\}}{\exp\{(\bar{x}\text{-}\mu_1)^2/(2\sigma_1^2+2\sigma^2/n)\}}\,\right]^{-1} \qquad (4)$$

It is easily checked that for any fixed $\bar{x}$ and p, the right hand side of (4) tends to one as σ_1^2 increases. Thus, for any *fixed prior* $p\{H_0\}=p$, the posterior probability of the null hypothesis can be made as close to one as desired, *whatever the data* for a sufficiently large prior variance $\sigma_1{}^2$. This is rather disturbing, for a large prior variance has been traditionally accepted as a description of vague initial knowledge.

A similar behaviour was found by Bartlett (1957) in his reply to Lindley's (1957) statistical paradox, and was later mentioned by Dempster (1971) in the Waterloo Symposium. They pointed out that, if a uniform distribution over a finite interval is chosen as $p\,(\mu|H_1)$, then $p\,(H_0|D)$ tends to one, whatever the data, as the size of the interval increases. It is clear that the same type of result will hold with any other distributional assumption for $p\,(\mu|H_1)$.

To us, this suggests that to obtain a sensible reference posterior probability for H_0, the value of p cannot be fixed and it is bound to depend on the form of $p(\mu|H_1)$. In the next Section, the method of maximizing the missing information is used to derive the reference posterior probability of a simple null hypothesis when the alternative is composite, and the precise form of this dependence is obtained. In Section 3, Lindley's paradox and the normal example are discussed in detail. In Section 4 asymptotic results are obtained for well-behaved probability models and finally, in Section 5, the main conclusions are reviewed and discussed.

2. A SOLUTION

Let ϵ be an experiment which produces some data D, the joint distribution of which $p\,(D|\theta)$ is indexed by some unknown parameter $\theta\epsilon\Theta$. Let H_0 be the simple null hypothesis that $\theta=\theta_0$. There is no posterior probability for H_0 without a mixed prior which allocates a positive amount of probability to H_0. Let this prior be

$$p\,(H_0) = p \qquad p\,(H_1) = 1\text{-}p \qquad (5)$$
$$p\,(\theta|H_1) = p\,(\theta)$$

For fixed $p\,(\theta)$ and p, the amount of *missing information* (Bernardo, 1979b), about θ, i.e. the maximum amount of information that ϵ may possibly supply about θ is defined to be

$$H(p) + (1\text{-}p)\, I^{\theta}\{\epsilon, p(\theta)\} \tag{6}$$

where $H(p) = -p\log p - (1\text{-}p)\log(1\text{-}p)$ is the entropy of the prior marginal distribution of H_0, and

$$I^{\theta}\{\epsilon, p(\theta)\} = \int\int p(\theta, D)\log\frac{p(\theta|D)}{p(\theta)}\, dD\, d\theta$$

is the amount of information about θ that, given H_1, one may expect from the experiment ϵ.

Here, and in the rest of the paper, we use the word information in the precise sense of Shannon (1948) and Lindley (1956). A decision-theoretic argument for the use of such a measure of information in scientific inference may be found in Bernardo (1979a).

Expression (6) has a simple intuitive interpretation; the maximum amount of information that, given the mixed prior distribution (5), the experiment ϵ may possibly be expected to supply, consists of the knowledge of whether H_0 is or is not true, which provides an amount of information $H(p)$, plus the amount of information about θ that ϵ may be expected to provide if H_1 is true, times the prior probability of H_1.

Taking in (6) derivatives with respect to p and equating to zero we find

$$\log\frac{1\text{-}p}{p} - I^{\theta}\{\epsilon, p(\theta)\} = 0$$

so that, for fixed $p(\theta)$, the prior probability π which maximizes the amount of missing information about θ is such that

$$\frac{1\text{-}\pi}{\pi} = \exp\,[I^{\theta}\{\epsilon, p(\theta)\}] \tag{7}$$

On the other hand, given the prior specification (5), the posterior probability of the null hypothesis is, using Bayes theorem,

$$p\,\{H_0|D\} = \left(1 + \frac{1\text{-}p}{p}\,\frac{p(D|H_1)}{p(D|H_0)}\right)^{-1}$$

so that, using (7), we obtain that for *fixed* $p(\theta)$, the *reference* posterior probability of the null hypothesis is

$$\pi\{H_0|D\} = \left(1 + \exp[I^{\theta}\{\epsilon, p(\theta)\}]\,\frac{\int p(D|\theta)\,p(\theta)\,d\theta}{p(D|\theta_0)}\right)^{-1} \tag{8}$$

To see how this works, consider again the normal example discussed before, so that $D = \{x_1, \ldots, x_n\}$ is a random sample from a normal distribution $N(x|\mu, \sigma^2)$, the null hypothesis is $\mu = \mu_0$, and the prior distribution of μ under H_1 is $p(\mu) = N(\mu|\mu_1, \sigma^2)$.

It is easily found that, in this case,

$$I^{\mu}\{\epsilon, p(\mu)\} = \tfrac{1}{2}\log\frac{\sigma_1^2 + \sigma^2/n}{\sigma^2/n}$$

so that, using (7)

$$\frac{1\text{-}\pi}{\pi} = \exp[I^{\mu}\{\epsilon, p(\mu)\}] = \left(\frac{\sigma_1^2 + \sigma^2/n}{\sigma^2/n}\right)^{1/2}$$

and, substituting into (4), one has

$$\pi\{H_0|D\} = \left(1 + \frac{\exp\{(\bar{x}\text{-}\mu_0)^2/(2\sigma^2/n)\}}{\exp\{(\bar{x}\text{-}\mu_1)^2/(2\sigma_1^2 + 2\sigma^2/n)\}}\right)^{-1} \tag{9}$$

which does *not* tend to one as σ_1^2 increases. To study the behavior of (9) let us define γ_x and γ_1 such that

$$\bar{x} = \mu_0 + \gamma_x\,\sigma/\sqrt{n}$$
$$\mu_1 = \mu_0 + \gamma_1\,\sigma_1$$

so that γ_x and γ_1 measure respectively, in standard units, how far the sample and the prior mean are from the null hypothesis.

Substitution into (9) yields

$$\pi\{(H_0|D)\} = \left(1 + \exp\{\tfrac{1}{2}(\gamma_x^2 - \frac{(\gamma_x\sigma/\sqrt{n}\text{-}\gamma_1\sigma_1)^2}{\sigma_1^2 + \sigma^2/n})\}\right)^{-1}$$

which, when either σ_1^2 or n increases, tends to

$$\pi\{H_0|D\} = [1 + \exp\{\tfrac{1}{2}(\gamma_x^2\text{-}\gamma_1^2)\}]^{-1} \tag{10}$$

or, in terms of odds,

$$\frac{\pi(H_1|D)}{\pi(H_0|D)} = \exp\{\tfrac{1}{2}(\gamma_x^2-\gamma_1^2)\} \qquad (11)$$

We believe this is a very reasonable result. It says essentially that the only important features are the distances in standard units, γ_x and γ_1 from the sample and from the prior mean to the null hypothesis. Under the null hypothesis, γ_x will be moderate, $(\gamma_x^2-\gamma_1^2)$ will not be too large, $\pi\{H_0|D\}$ will never be too close to zero, and we shall *not* reject H_0. Under the alternative hypothesis, γ_x will increase as $\sqrt{n}$, $(\gamma_x^2-\gamma_1^2)$ will increase as n, $\pi\{H_0|D\}$ will tend to zero and we shall eventually reject the false null.

3. LINDLEY'S PARADOX

To compare our results with those obtained by classical hypothesis testing one has to use a 'non-informative' or reference prior for $p(\theta|H_1)$ and not just for $p(H_0)$. Maximizing the missing information given H_1 gives, under regularity conditions, (see Bernardo, 1979b) Jeffreys' prior, i.e.

$$\pi(\theta|H_1) = \pi(\theta) = \frac{i(\theta)^{1/2}}{\int_\Theta i(\theta)^{1/2}\,d\theta} \qquad (12)$$

where, if necessary, θ has been restricted to a set of finite measure for the integral in (12) to exist.

In the normal example discussed earlier, $i(\mu)^{1/2} = 1/\sigma$ so that, using (12), if $\mu\in]-A,A[$,

$$\pi(\mu|H_1) = \pi(\mu) = \frac{1}{2A}$$

With this prior, if $p(H_0) = p$, the posterior probability of H_0 is

$$p(H_0|D) = \left(1 + \frac{1-p}{p}\,\frac{1/(2A)}{(n/2\pi\sigma^2)^{1/2}\exp\{-n(\bar{x}-\mu_0)^2/2\sigma^2\}}\right)^{-1} \qquad (13)$$

which, for fixed p, tends to one as A increases. However,

$$I^\mu\{\epsilon,\pi(\mu)\} = \int \pi(\mu)\int p(\bar{x}|\mu)\log\frac{p(\bar{x}|\mu)}{\int p(\bar{x}|\mu)\,\pi(\mu)\,d\mu}\,d\bar{x}\,d\mu$$

$$= \tfrac{1}{2}\log\frac{n}{2\pi e\sigma^2} + \log 2A$$

so that, using (7)

$$\frac{1-\pi}{\pi} = \exp\,[I^\mu\{\epsilon,\pi(\mu)\}] = \left(\frac{n}{2\pi e\sigma^2}\right)^{1/2} 2A \qquad (14)$$

and substituting into (13)

$$\pi(H_0|D) = \left(1 + \exp\{\tfrac{1}{2}[(n/\sigma^2)(\bar{x}-\mu_0)^2 - 1]\}\right)^{-1} \qquad (15)$$

which does *not* depend on the arbitrary constant A. If γ_x is defined such that $\bar{x} = \mu_0+\gamma_x\sigma/\sqrt{n}$, the last equation may be rewriten as

$$\pi(H_0|D) = [1+\exp\{\tfrac{1}{2}(\gamma_x^2-1)\}]^{-1} \qquad (16)$$

or, in terms of odds,

$$\frac{\pi(H_1|D)}{\pi(H_0|D)} = \exp\{\tfrac{1}{2}(\gamma_x^2-1)\} \qquad (17)$$

This establishes a one-to-one relationship between a significance level and a reference posterior probability for the null hypothesis. Thus, a result significant at 0.05 level, implies $\gamma_x = 1.96$ and therefore, using (17), odds of about 4 to 1 against the null hypothesis. In Table 1, the precise equivalences are given for a number of commonly used significance levels.

α	γ_x	$\pi(H_1\|D)/\pi(H_0\|D)$	$\pi(H_0\|D)$
0.1	1.645	2.347	0.299
0.05	1.960	4.140	0.195
0.01	2.576	16.741	0.0564
0.005	2.807	31.175	0.0310
0.001	3.291	136.366	0.00728
0.0001	3.891	1176.078	0.000850
0.00001	4.417	10456.135	0.0000956

TABLE 1

EQUIVALENCE BETWEEN SIGNIFICANCE LEVELS AND REFERENCE POSTERIOR PROBABILITIES FOR THE SIMPLE NORMAL CASE

The expected value $\gamma_x^2 = (\bar{x}-\mu_0)^2\,n/\sigma^2$ under the null hypothesis is one. Thus, under the null hypothesis, $\pi(H_0|D)$ will tend to 1/2 as n increases. Under the alternative hypothesis, the expected value of γ_x^2 is $1 + n\left(\frac{\mu-\mu_0}{\sigma}\right)^2$

and thus $\pi(H_0|D)$ will tend to 0 as n increases. This is precisely the sort of behaviour one could expect: one can never prove the null hypothesis to be true but one may reject it when it is false. This is illustrated in Figure 1, where the value of $\pi(H_0|D) = \pi(\mu=0|x_1,x_2,\ldots,x_n)$ given by (15) is plotted as a function of the sample size n for a sequence of data simulated from (i) $N(x|0,1)$, (ii) $N(x|1,1)$ and (iii) $N(x|3,1)$. As expected, the reference posterior (i) oscillate around 0.5 while those of (ii) and (iii) tend to zero, that of (iii) much more rapidly than that of (ii).

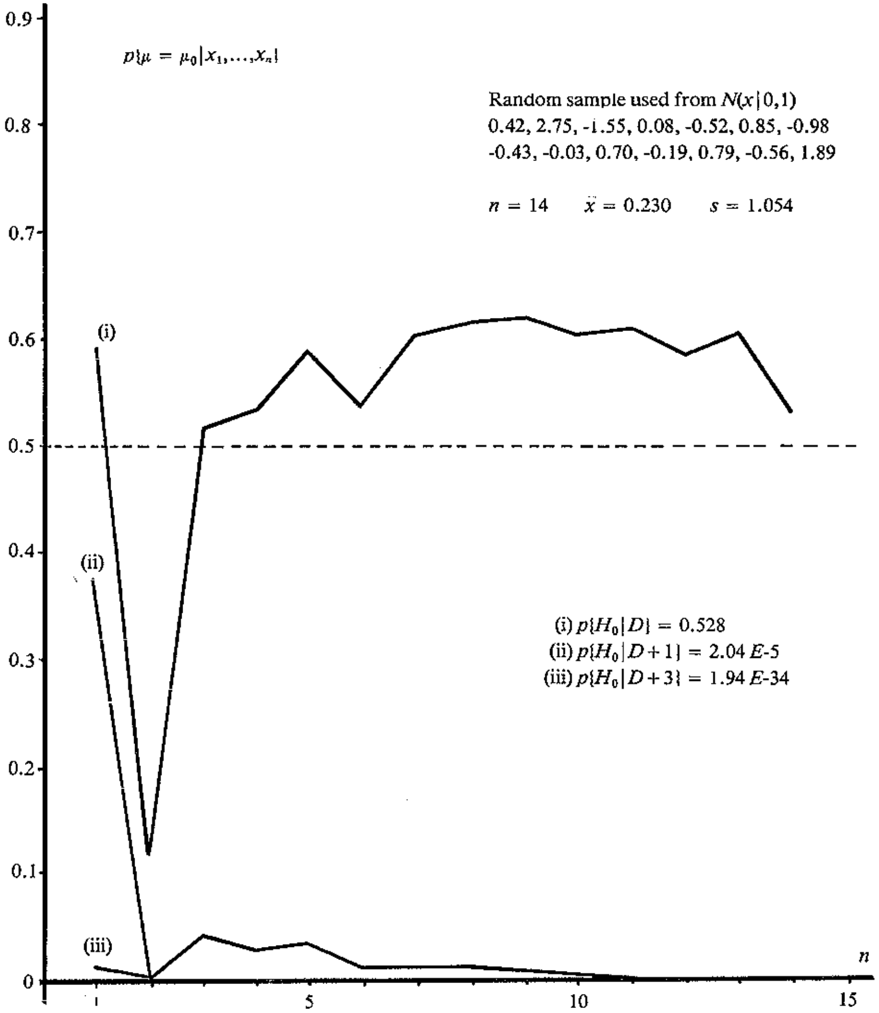


FIGURE 1

POSTERIOR PROBABILITY OF THE NULL HYPOTEHSIS $\mu = \mu_0$ FOR SIMULATED DATA FROM (i) $N(x|0,1)$, (ii) $N(x|1,1)$, (iii) $N(x|3,1)$

This results seems to clarify Lindley's (1957) paradox. In his analysis, Lindley does not introduce the constant $2A$. Although he acknowledges that this is done by mistake in his reply to Bartlett's comments, this is a step in the right direction, for this effectively means to choose $(1\text{-}p)/p = 2A$. However, he does not introduce the factor $\sqrt{n}$ in (14) which is necessary, from an information-theoretical point of view, to compensate for the different dimensionalities of H_0 and H_1. This missing factor $\sqrt{n}$ is responsible for wide discrepancy he finds between the classical approach and his solution. Indeed, if $\bar{x}$ is found significant at say, 5%, then $\gamma_x = 1.96$ and we find for H_0 a reference posterior probability 0.195 *independent* of n. Lindley, on the other hand, finds that his posterior probability of H_0, for fixed γ_x, increases with n and tends to one as n tends to infinity.

If our analysis is correct, γ_x is indeed sufficient to decide whether one has strong evidence against the null hypothesis and (16) may be used to translate significance levels into posterior probabilities.

4. ASYMPTOTIC RESULTS

The exact value of $I^{\theta}\{\epsilon,p(\theta)\}$, which according to (7) is needed to obtain the reference prior for H_0, is often difficult to obtain. Good approximations are available however when the data set D produced by ϵ is large.

Indeed, if $D=\{x_1,x_2,\ldots,x_n\}$, and n is large, one has (Stone, 1958; Ibragimov & Hasminsky, 1973), that

$$I^{\theta}\{\epsilon,p(\theta)\} = \tfrac{1}{2}\log\frac{n}{2\pi e} + \int p(\theta)\log\frac{i(\theta)^{1/2}}{p(\theta)}\,d\theta + o(1)$$

and therefore if, using (16), $\pi(\theta) = i(\theta)^{1/2}/\int_{\Theta} i(\theta)^{1/2}\,d\theta$,

$$I^{\theta}\{\epsilon,p(\theta)\} = \tfrac{1}{2}\log\frac{n}{2\pi e} + \log\{\int_{\Theta} i(\theta)^{1/2}\,d\theta\}$$

so that, using (7)

$$\frac{1\text{-}\pi}{\pi} = \left(\frac{n}{2\pi e}\right)^{1/2}\int_{\Theta} i(\theta)^{1/2}\,d\theta \qquad (18)$$

On the other hand, the posterior probability of $\theta = \theta_0$ with mixed prior structure such as (5) is

$$p(H_0|D) = \left(1 + \frac{1\text{-}p}{p}\,\frac{\int p(D|\theta)p(\theta)\,d\theta}{p(D|\theta_0)}\right)^{-1} \qquad (19)$$

For large n, the maximum likelihood estimate of θ, $\hat{\theta}$, will be sufficient and may replace D. Moreover,

$$p(\hat{\theta}\,|\,\theta) \cong N(\hat{\theta}\,|\,\theta,\ 1/ni(\theta))$$

so that

$$p(\hat{\theta}) \cong \int N\left(\hat{\theta}\,|\,\theta,\ \frac{1}{ni(\theta)}\right) \frac{i(\theta)^{1/2}}{\int_{\Theta} i(\theta)^{1/2} d\theta}\, d\theta \cong \frac{i(\hat{\theta})^{1/2}}{\int_{\Theta} i(\theta)^{1/2} d\theta}$$

Using (18) into (19) one obtains

$$\pi(H_0|D) = \left(1 + \left(\frac{i(\hat{\theta})}{i(\theta_0)}\right)^{1/2} \exp\{\tfrac{1}{2}[ni(\theta_0)(\hat{\theta}-\theta_0)^2 - 1]\}\right)^{-1}$$

Finally, if γ_x is defined such that

$$\hat{\theta} = \theta_0 + \gamma_x/\sqrt{ni(\theta_0)}$$

one has

$$\pi(H_0|D) = \left(1 + \left(\frac{i(\hat{\theta})}{i(\theta_0)}\right)^{1/2} \exp\{\tfrac{1}{2}(\gamma_x^2 - 1)\}\right)^{-1} \tag{20}$$

or, in terms of odds,

$$\frac{\pi(H_1|D)}{\pi(H_0|D)} = \left(\frac{i(\hat{\theta})}{i(\theta_0)}\right)^{1/2} \exp\{\tfrac{1}{2}(\gamma_x^2 - 1)\} \tag{21}$$

which has, of course, the same qualitative behaviour as the equations (10) and (17) of the exact normal case.

Consider, for instance, that the observation of n binomial trials have produced r succeses and we want to test whether $\theta = \theta_0$. It is easily verified that in this case $i(\theta) = \{\theta(1-\theta)\}^{-1}$ so that the odds against the null hypothesis will approximately be, using (21),

$$\frac{\pi(H_1|D)}{\pi(H_0|D)} = \left(\frac{\hat{\theta}(1-\hat{\theta})}{\theta_0(1-\theta_0}\right)^{1/2} \exp\{\tfrac{1}{2} \frac{n}{\theta_0(1-\theta_0)} (\hat{\theta}-\theta_0)^2 - \tfrac{1}{2}\}$$

where $\hat{\theta} = r/n$. In Table 2, the values of $\pi(H_1|D)/\pi(H_0|D)$ and those of $\pi(H_0|D)$ are given when $\theta_0 = 0.2$ and $n = 1000$ for different values of r.

r	$\pi(H_1\|D)/\pi(H_0\|D)$	$\pi(H_0\|D)$
200	0.607	0.622
190	0.674	0.597
210	0.700	0.588
180	0.960	0.510
220	1.036	0.491
160	4.107	0.196
240	4.785	0.172
140	47.362	0.021
260	59.871	0.016
120	1468	6.8×10^{-4}
280	2029	4.9×10^{-4}
100	122066	8.2×10^{-6}
300	186459	5.3×10^{-6}

TABLE 2

REFERENCE PROBABILITIES FOR THE NULL HYPOTHESIS $\theta = 0.2$ WHEN $n = 1000$, FOR DIFFERENT VALUES OF r

It is easy to check that these results are consistent with those of Table 1 using the fact that, under the null hypothesis, the standard deviation or r is about 20. As one would expect from the situation of the null hypothesis within the interval (0,1), deviations to the right of θ_0 provide more evidence against H_0 than deviations of the same size to the left of θ_0.

In the Soal and Bateman (1954) experiment to test the telepathic powers of Mrs Stewart, mentioned by Lindley (1957), the null hypothesis (no telephatic powers) is $\theta_o = 0.2$, and 9410 succeses where obtained out of 37100 trials. The posterior probability of the null hypothesis results to be $p(H_0|D) \cong 10^{-144.65}$; the evidence for Mrs Stewart telepathic powers is rather strong. Indeed, we find, about 45.000 times stronger than Lindley suggested.

5. DISCUSSION

We mentioned in Section one, that derivation of reference posterior probabilities for the null hypothesis does not present problems if the null and the alternative have the same dimension. It may be argued that these are the only interesting cases and that a simple null versus a composite alternative is a mathematical abstraction. We believe however that the problem is worthy of investigation. Indeed, (i) a scientific theory may imply a precise value for a given magnitude; to check whether the data are compatible with the theory is precisely to test this null hypothesis, (ii) even if one is really only interested in whether or not θ belongs to a small neighbourghood of θ_0, testing $\theta = \theta_0$ gives a reasonable approximate answer and indicates, through the reference posterior distribution given $H_1, \pi(\theta | H_1, D) \propto p(D | \theta)\pi(\theta)$ and indication of possible alternative values of θ if θ_0 is rejected.

On the other hand, scientists find often natural to frame their research in terms of cheking whether the data observed falsify or not, in statistical terms, a particular theory. Moreover, they have been doing so for years, rather successfully, using classical hypothesis testing. It is natural to enquire whether an explanation may be given from a Bayesian point of view. If the procedures developed here are accepted, classical hypothesis testing might not be too bad, *provided* that no confidence meaning is attached to the significance level. Our approach establishes a correspondance, for each particular problem, between significance levels and reference posterior probabilities. It also implies that the ubiquituous 0.05 significance level only suggests evidence of about 4 to 1 against the null when, in most applications, a larger amount of evidence against the null would be required before rejection.

The main results of our approach are that (i) one can 'reject' a null hypothesis, i.e. $\pi(H_0 | D)$ may approach zero, but one cannot 'prove' it, i.e. $\pi(H_0 | D)$ never approaches one, and (ii) the possible evidence against the null hypothesis is roughly summarized in the standarized distance γ_x between the null value and the likelihood estimate of the parameter.

Both (i) and (ii) have been traditionally accepted in practice, and also make sense intuitively from a Bayesian viewpoint. A foundations type argument for (i) has been given by Popper (1958,ch.10). Another argument for (ii) may be given; different authors, Dempster (1971) for instance, has suggested looking at the tails of the posterior distribution of θ defined by θ_0 in order to 'test' the null. It easy to see that those tails approximately depend on the data, only through the standarized distance $\gamma_x = \sqrt{\{ni(\theta_0)\}}(\theta - \theta_0)$.

The reference posterior for the null given by (20) provides the relationship between γ_x and $\pi(H_0 | D)$ for large samples. If $\hat{\theta}$ is close to θ_0, the factor $\{i(\hat{\theta})/i(\theta_0)\}^{1/2}$ will be close to one; if not, γ_x will be large and that factor will be dominated by the exponential. Thus, the simple relation

$$\pi(H_0 | D) \cong [1 + \exp\{\tfrac{1}{2}(\gamma_x^2 - 1)\}]^{-1}$$

will approximately be true for large samples of any model. The results obtained are

$\|\gamma_x\|$	0	1	2	3	4	5
$\pi(H_1 \| D)/\pi(H_0 \| D)$	0.606	1.000	4.482	54.60	1808	1.6×10^5
$\pi(H_0 \| D)$	0.622	0.500	0.182	0.018	5.5×10^{-4}	6.1×10^{-6}

We have used a particular definition of information and a rather specific way of handling this in order to obtain our reference posteriors. From a theoretical point of view, the choice of a logarithmic measure of information may be axiomatically defended (Good, 1966; Bernardo 1979a) and its properties seem to be adequate (Lindley 1956). Moreover, the procedure of maximizing the missing information has been proved capable of unifying previous results on 'non-informative' priors and has been shown to produce sensible results to some controversial problems (Bernardo, 1979b). Finally, any argument on the foundations of a procedure should take account of the results it gives rise. We find that our results are intuitively reasonable and that they provide a Bayesian interpretation of classical hypothesis testing. Whether or not this view may be shared by others will shortly become apparent in the discussion.

REFERENCES

BARTLETT, M.S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika* **44**, 533-534.

BERNARDO, J.M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686-690.

— (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113-147 (with discussion).

DEMSTER, A.P. (1971). Model searching and estimation in the logic of inference. In *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), 56-81. Toronto: Holt, Rinehart and Winston.

GOOD, I.J. (1966). A derivation of the probabilistic explication of information. *J. Roy. Statist. Soc. B* **28**, 578-581.

IBRAGIMOV, I.A. and HASMINSKY, R.Z. (1973). On the information in a sample about a parame ter. *2nd Internat. Symp. Information Theory.* (Petrov and Csaki eds.) 295-309. Budapest: Akedemiai Kiado.

JEFFREYS, H. (1939/67). *Theory of probability*. Oxford: Clarendon Press.

LINDLEY, D.V. (1956). On a measure of the information provided by an experiment. *Ann. Math.* **27**, 986-1005.

— (1957). A statistical paradox. *Biometrika* **44**, 187-192.

POPPER, K.R. (1958). *The logic of scientific discovery*. London: Hutchinson and Co.

SHANNON, C.E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 379-423, 623-656.

SOAL, S.G. and BATEMAN, F. (1954). *Modern Experiments in Telepathy*. London: Faber and Faber.

STONE, M. (1958). *Studies with a measure of information*. Ph. D. Thesis, University of Cambridge.

DISCUSSION

E.T. JAYNES (*Washington University*):

It is always interesting to recall the arguments that Jeffreys used to find priors. The case recounted by Zellner is a typical example where it appears at first glance that we have nothing to go on; yet by thinking more deeply, Jeffreys finds something. He shows an uncanny ability to see intuitively the right thing to do, although the rationalization he offers is sometimes, as Laplace said of Bayes' argument, "fine et très ingénieuse, quoiqu'un peu embarrassée" It was from studying these flashes of intuition in Jeffreys that I become convinced that there must exist a general formal theory of determination of priors by logical analysis of the prior information-and that to develop it is today the top priority research problem of Bayesian theory.

Pragmatically, the actual results of the Jeffreys-Zellner-Siow and Bernardo tests seem quite reasonable; without considerable analysis one could hardly say how or whether we should want them any different. Likewise, there is little to say about the mathematics, since once the premises are accepted, all else seems to follow in a rather straightforward and inevitable way. So let us concentrate on the premises; more specifically, on the technical problems encountered in both works, caused by putting that lump of prior probability on a single point $\lambda = 0$.

1. *The problem*

In most Bayesian calculations the same prior appears in numerator and denominator, and any normalization constant cancels out. Usually, passage to the limit of an "uninformative" improper prior is then uneventful; i.e., our conclusions are very robust with respect to the exact prior range. But in Jeffreys' significance test this robustness is lost, since $K = p(D|H_o)/p(D|H_1)$ contains in the denominator an uncancelled factor which is essentially the prior density $\pi(\lambda)$ at $\lambda = \bar{x}$. Then in the limit of an improper prior we have $K \to \infty$ independently of the data D, a result given by Jeffreys (1939, p. 194, Eq. 10), and since rediscovered many times. Note that the difficulty is not due solely to the different dimensionality of the parameter spaces; it would appear in any problem where we think of H_o as specifying a definitive, fixed prior range, but fail to do the same for H_1.

Jeffreys (1961) dealt with this and other problems by using a Cauchy prior $\pi(\lambda|\sigma)$ scaled on σ in the significance test, although he would have used a uniform prior $\pi(\lambda) = 1$ in the same model H_1 had he been estimating λ. But then a question of principle rears up. To paraphrase Lindley's rhetorical question: Why should our prior knowledge, or ignorance, of λ depend on the question we are asking about it? Even more puzzling: why should it depend on another parameter σ, which is itself unknown? One feels the need for a clearer rationalization.

Furthermore, the difficulty was not really removed, but only concealed from view, by Jeffreys' procedure. All his stated conditions on the prior would have been met equally well had he chosen a Cauchy distribution with interquartile span 4σ instead of σ; but then all his K-values would have been quadrupled, leading to indifference at a very different value of the t-statistic [see Eq. (5-13) below]. We do not argue that Jeffreys made a bad choice; quite the contrary. Our point is rather that in his choice there were elements of arbitrariness, arising from a still unresolved question of principle. Pending that resolution, one is not in a position to say much about the "uniqueness" or "objectivity" of the test beyond the admitted virtue of yielding results that seem reasonable.

Bernardo comes up against just the same problem, but deals with it more forthrightly. Finding again that the posterior probability P_o of the null hypothesis H_o increases with the prior variance σ_1 in a disconcerting way, he takes what I should describe as a meat-axe approach to the difficulty, and simply chops away at its prior probability p until $P_o = pk/(pk + 1\text{-}p)$ is reduced to what he considers reasonable (from the Jeffreys-Zellner-Siow standpoint he chops a bit too much, since his P_o tends only to 1/2 on prolonged sampling when H_o is true). This approach has one great virtue: whereas the Jeffreys results tended to be analytically messy, calling for tedious approximations, Bernardo emerges triumphantly (in the limit of large σ_1) with a beautifully neat expression (Eq. (11)) which has also, intuitively, a clear ring of truth to it.

But for this nice result, Bernardo pays a terrible price in unBayesianity. He gets it only by making p vary with the sample size n, calling for another obvious paraphrase of Lindley. This elastic quality of his prior is rationalized by an information-theoretic argument; it is, in a sense, the prior for which one would expect (before seeing the data) to learn the most from the experiment. But is this the property one wants?

If a prior is to incorporate the *prior information* we had about λ before the sample was observed, it cannot depend on the sample. The difficulty is particularly acute if the test is conducted sequentially; must we go back to the beginning and revise our prior as each new data point comes in? Yet after all criticisims I like the general tone of Bernardo's result, and deplore only his method of deriving it.

The common plot of these two scenarios is: we (1) start to apply Bayes' theorem in what seems a straightforward way; (2) discover that the result has an unexpected dependence on the prior; (3) patch things up by tampering with the prior until the expected kind of result emerges. The Jeffreys and Bernardo tamperings are similar in effect, although they offer very different rationalizations for what they do. But in both cases the tampering has a mathematical awkwardness and the rationalization a certain contrived quality, that leads one to ask whether some important point has been missed.

Now, why should that first result have been unexpected? If, according to H_1, we know initially only that λ is in some very wide range $2\sigma_1$, and we then receive data showing that it is actually within $\pm\sigma/\sqrt{n}$ of the value predicted by H_o, -as a physicist would put it, "the data agree with H_o to within experimental error"- that is indeed very strong evidence in favor of H_o. Such data ought to yield a likelihood ratio $K = \sqrt{n}\sigma_1/\sigma$

increasing with σ_1, just as Bernardo finds. This first result is clearly the correct answer to the question Q_1 that was being asked.

If we find that answer disconcerting, it can be only because we had in the back of our minds a different, unenunciated question Q_2. On this view, the tampering is seen as a mutilation of equations originally designed to answer Q_1, so as to force them to answer instead Q_2.

The higher-level question: "Which question should we ask?" does not seem to have been studied explicitly in statistics, but from the way it arises here, one may suspect that the answer is part of the necessary "software" required for proper use of Bayesian theory. That is, just as a computer stands ready to perform any calculation we ask of it, our present theory of Bayesian inference stands ready to answer any question we put to it. In both cases, the machine needs to be programmed to tell it which task to perform. So let us digress with some general remarks on question-choosing.

2. *Logic of Questions*

For many years I have called attention to the work on foundations of probability theory by R.T. Cox (1946,1961) which in my view provides the most fundamental and elegant basis for Bayesian theory. We are familiar with the Aristotelian deductive logic of propositions; two propositions are equivalent if they say the same thing, from a given set of them one can construct new propositions by conjunction, disjunction, etc. The probability theory of Bernouilli and Laplace included Aristotelian logic as a limiting form, but was a mathematical extension to the intermediate region $(0<p<1)$ between proof and disproof where, of necessity, virtually all our actual reasoning takes place. While orthodox doctrine was rejecting this as arbitrary, Cox proved that it is the only consistent extension of logic in which degrees of plausibility are represented by real numbers.

Now we have a new work by Cox (1978) which may prove to be of even more fundamental importance for statistical theory. Felix Klein (1939) suggested that questions, like propositions, might be used as logical elements. Cox shows that in fact there is an exactly parallel logic of questions: two questions are equivalent if they ask the same thing, from a given set of them one can construct new questions by conjunction (ask both), disjunction (ask either), etc. All the "Boolean algebra" of propositions may be taken over into a new symbolic algebra of questions. Every theorem of logic about the "truth value" of propositions has a dual theorem about the "asking value" of questions.

Presumably, then, besides our present Bayesian statistics -a formal theory of optimal inference telling us which propositions are most plausible- there should exist a parallel formal theory of optimal inquiry, telling us which questions are most informative. Cox makes a start in this direction, showing that a given question may be defined in many ways by the set of its possible answers, but the question possesses an entropy independent of its defining set, and the entropies of different questions obey algebraic rules of combination much like those obeyed by the probabilities of propositions.

The importance of such a theory, further developed, for the design of experiments and the choosing of procedures for inference, is clear. For over a century we have argued over which *ad hoc* statistical procedures ought to be used, not on grounds of any demonstrable properties, but from nothing more than ideological committments to various preconceived positions. There is still a great deal of this in my exchanges with Margaret Maxfield and Oscar Kempthorne in Jaynes (1976), and even a little in the exchange with Dawid, Stone, and Zidek over marginalization in Jaynes (1980). A formal theory of optimal inquiry might resolve differences of opinion in a way that Wald-type decision theory and Shannon-type information theory have not accomplished.

Our present problem involves a special case of this. If, seeing the answer to question Q_1 we are unhappy with it, what alternative question Q_2 did we have, unconsciously, in the back of our minds? Is there a question Q_3 that is the optimal one to ask for the purpose at hand? Since the conjectured formal theory of inquiry is still largely undeveloped, we try to guess some of its eventual features by studying this example.

Note that the issue is not which question is "correct". We are free to ask of the Bayesian formalism any question we please, and it will always give us the best answer it can, based on the information we have put into it. But still, we are in somewhat the position of a lawyer at a courtroom trial. Even when he has on the stand a witness who knows all the facts of the case and is sworn to tell the truth, the information he can actually elicit from this witness still depends on his adroitness in asking the right questions.

If his witness is unfriendly, he will not extract any information at all unless he knows the right questions to force it out, phrasing them as sharp leading questions and demanding unequivocal "yes" or "no" answers. But if a witness is friendly and intelligent, one can get all the information desired more quickly by asking simply, "Please tell us in your own words what you know about the case?" Indeed, this may bring out unexpected new facts for which one could not have formulated any specific question.

Significance tests which specify a sharply defined hypothesis and preassigned significance level, and demand to know whether the hypothesis does or does not pass at that level, therefore in effect treat probability theory as an unfriendly witness and automatically preclude any possibility of getting more information than that one bit demanded.

Suppose we try instead the opposite tactic, and regard Bayesian formalism as a friendly witness, ready and willing to give us all the pertinent information in our problem even information that we had not realized was pertinent if we only allow if the freedom to do so. Instead of demanding the posterior probability of some sharply formulated null hypothesis H_0, suppose we ask of it only, "Please tell us in your own words what you know about λ?" Perhaps by asking a less sharp and restrictive question, we shall elicit more information.

3. *Information from questions*

Evidently, to deal with such problems one ought to be an information theorist, and not only in the narrow sense of One-Who-Uses-Entropy. In the present problem we are concerned not only with the range of possible answers, as measured by the entropy of a question, but also with the specific kind information that the question can elicit. In the following we use the word "information" in this semantic sense rather than the entropy sense.

All statistical procedures are in the last analysis prescriptions for information processing: what information have we put into our mathematical machine, and what information are we trying to get out of it? In these terms, what is the difference -if any- between significance testing and estimation? Having put certain information (model, prior, and data) into our hopper, we may carry out either, by asking different questions. But the answers to different questions do not necessarily convey different information.

The tests considered by Zellner and Bernardo sought information that can help us decide whether to adopt a new hypothesis H_1 with a value of λ different from its currently supposed value $\lambda=0$. Presumably, any procedure which yields the same information would be equally acceptable for this purpose, even though current pedagogy might not call it a "significance test"

Now this information criterion establishes an ordering of different procedures, or "tests", rather like the notion of admissibility. If test B (which answers question Q_B) always gives us the same information as test A, and sometimes more, then B may be said to dominate A in the sense of information yield, or question Q_B dominates Q_A in "asking power"; and if B requires no more computation, on what grounds could one ever prefer A?

In my work of 1976 (p. 185 and p. 219), I showed that the original Bayesian sigificance test of Laplace, which asks for the posterior probability P_1 of a one-sided alternative hypothesis, dominates the traditional orthodox t-test and F-test in just this sense. That is, given P_1 we know what the verdict would be, at any significance level, for all three of the corresponding orthodox tests (one equal-tails and two one-sided; but the veredict of any one orthodox test is far from determining P_1. Thanks to Cox, we have now a much broader view of this phenomenon.

Let us call a question *simple* if its answer is a single real number; or in Cox's terminology, if its irreducible defining set is a set of real numbers. For example: "What is the probability that λ, or some function of λ, lies in a certain region R?"

In any problem involving a single parameter λ for which there is a single sufficient statistic u, then given any simple question Q_A about λ, the answer will be, necessarily, some function $a(u)$. Given any two such questions Q_A, Q_B and any fixed prior information, the answers $a(u)$, $b(u)$, being functions of a single variable u, must obey some functional relation $a=f(b)$. If $f(b)$ is single-valued, then the answer to Q_B tells us everything that the answer to Q_A does. As Cox puts it, "An assertion answering a question answers every implicate of that question". If the inverse function $b=f^{-1}(a)$ is not single-valued, then Q_B dominates Q_A.

In the case of a single sufficient statistic, then, any simple question whose answer is a strict monotonic function of u, yields all the information that we can elicit about λ, whatever question we ask; and it dominates any simple question whose answer is not a strict monotonic function of u. But this is just the case discussed by Bernardo; he considers σ known, and consequently $\bar{x}$ is sufficient statistic for λ. Since his odds ratio $K(\bar{x})$ is not a strict monotonic function of $\bar{x}$, we know at once that Bernardo's test is dominated by another.

The Jeffreys-Zellner-Siow tests are more subtle in this respect, since σ is unknown, and consequently there are two jointly sufficient statistics $(\bar{x},s)$. Given two simple questions Q_A,Q_B with answers $a(\bar{x},s)$, $b(\bar{x},s)$, the condition that they ask essentially the same thing, leading to a functional relation $a=f(b)$, is that the Jacobian $J = \partial(a,b)/\partial(\bar{x},s)$ should vanish. If $J\neq 0$, then neither questions can dominate the other and no simple question can dominate both. But any two simple questions for which $(\bar{x},s)$ are uniquely recoverable as single-valued functions $\bar{x}(a,b)$, $s(a,b)$ will jointly elicit all the information that any question can yield, and thus their conjunction dominates any simple question.

We may, therefore, conclude the following. Since Jeffreys' test asks a simple question, whose answer is the odds ratio $K(\bar{x},s)$, it can be dominated by a compound question, the conjuction of two simple questions. Indeed, since K depends only on the magnitude of the statistic t, it is clear that Jeffreys' question is dominated by any one simple question whose answer is a strict monotonic function of t.

These properties generalize effortlessly to higher dimensions and arbitrary sets. Whenever sufficient statistics exist, the most searching questions for any statistical procedure, -whatever current pedagogy may call it- are those (simple or compound) from whose answers the sufficient statistics may be recovered; and all such questions elicit just the same information from the data.

As soon as I realized this, it struck me that this is exactly the kind of result that Fisher would have considered intuitively obvious from the start; however, a search of his collected works failed to locate any passage where such an idea is stated. Perhaps others may recall instances where he made similar remarks in private conversation; it is difficult to believe that he was unaware of it.

With these things in mind, let us re-examine the rationale of the Jeffreys-Zellner-Siow and Bernardo tests.

4. *What is our rationale?*

In pondering this -trying to see where we have confused two different questions and what the question Q_2 is- I was struck by the constrast between the reasoning used in the proposed tests and the reasoning that physicists use, in everyday practice, to decide such matters. We cite one case history; recent memory would yield a dozen equally good, which make the same point.

In 1958, Cocconi and Salpeter proposed a new theory H_1 of gravitation, which predicted that the inertial mass of a body is a tensor. That is, instead of Newton's $F=Ma$, one had $F_i=\Sigma M_{ij}a_j$. For terrestrial mechanics the principal axes of this tensor would be determined by the distribution of mass in our galaxy, such that with the x-axis directed toward the galactic center, $M_{xx}/M_{yy}=M_{xx}/M_{zz}=(1+\lambda)$. From the approximately known galactic mass and size, one could estimate (Weisskopf, 1961) a value $\lambda \cong 10^{-8}$

Such a small effect would not have been noticed before, but when the new hypothesis H_1 was brought forth it became a kind of challenge to experimental physicists: devise an experiment to detect this effect, if it exists, with the greatest possible sensitivity. Fortunately, the newly discovered Mössbauer effect provided a test with sensitivity far beyond one's wildest dreams. The experimental verdict (Sherwin, *et.al,* 1960) was that λ, if it exists, cannot be greater than $|\lambda| < 10^{-15}$. So we forgot about H_1 and retained our null hypothesis: H_o = *Einstein's theory of gravitation, in* which $\lambda = 0$.

From this and other case histories in which other conclusions were drawn, we can summarize the procedure of the physicist's significance test as follows: (A) Assume the alternative H_1, which contains a new parameter λ, true as a working hypothesis. (B) On this basis, devise an experiment which can measure λ *with the greatest possible* precision. (C) Do the experiment. (D) Analyze the data as a pure estimation problem-Bayesian, orthodox, or still more informal, but in any event leading to a final "best" estimate and a statement of the accuracy claimed: $(\lambda)_{est} = \lambda' \pm \delta\lambda$. It is considered good form to claim an accuracy δλ corresponding to at least two, preferably three, standard deviations. (E) Let λ_o be the correct value according to *the null hypothesis* H_o (we supposed $\lambda_o = 0$ above, but it is now best to bring it explicitly into view), and define the "statistic" $t \equiv (\lambda' - \lambda_o)/\delta\lambda$. Then there are three possible outcomes:

If $\lvert t\rvert < 1$, retain H_o,	STATUS QUO
If $\lvert t\rvert >> 1$, accept H_1,	AWARD NOBEL PRIZES
If $1 < \lvert t\rvert < 3$, withhold judgment	SEEK BETTER EXPERIMENTS

That is, to within the usual poetic license, the reasoning format in which the progress of physics takes place.

You see why I like the actual results reported here by Zellner and Bernardo, although I find their rationalizations puzzling. They did indeed find, *as the criterion for* accepting H_1, that the estimated deviation $|\lambda' - \lambda_o|$ should be large compared to the accuracy of the measurement, considered known $(\sigma/\sqrt{n})$ in Bernardo's problem, and estimated from the data in the usual way $(s/\sqrt{n})$ in Zellner's.

It is in the criterion for retaining H_o that we seem to differ; contrast the physicist's rationale with that usually advanced by statisticians, Bayesian or otherwise. When we retain the null hypothesis, our reason is not that it has emerged from the test with a high posterior probability, or even that it has accounted well for the data. H_o is retained for the totally different reason that if the most sensitive available test fails to detect its existence, the new effect $(\lambda - \lambda_o)$ can have no observable consequences. That is, we are still free to adopt the alternative H_1 if we wish to; but then we shall be obliged to use a value of λ so close to the previous λ_o that all our resulting predictive distributions will be indistinguishable from those based on H_o.

In short, our rationale is not probabilistic at all, but simply pragmatic; having nothing to gain in predictive power by switching to the more complicated hypothesis H_1, we emulate Ockham. Note that the force of this argument would be in no way diminished even if H_o had emerged from some significance test with an extremely low posterior probability; we would still have nothing to gain by switching. Our acceptance of H_1 when $|t| >> 1$ does, however, have a probabilistic basis, as we shall see presently.

Today, most physicists have never heard the term "significance test". Nevertheless, the procedure just described derives historically from the original tests devised by Laplace in the 18'th Century, to decide whether observational data indicate the existence of new systematic effects. Indeed, the need for such tests in astronomy was the reason why the young Pierre Simon developed an interest in probability theory, forty-five years before he became the *Marquis de Laplace.* This problem is therefore the original one, out of which "Bayesian statistics" grew.

As noted also by E.C. Molina (1963) in introducing the photographic reproduction of Bayes' paper, even the result that we call today "Bayes' theorem" was actually given not by Bayes but by Laplace (the only valid reason I have found for calling it "Bayes' theorem" was provided at this meeting; "There's no theorem like Laplace's theorem" does not set well to Irving Berlin's music). Molina also offers some penetrating remarks about Boole's work, showing that those who have quoted Boole in support of their criticisms of Bayes and Laplace may have mistaken Boole's intention.

Now, although Laplace's tests were thoroughly "Bayesian" in the sense just elucidated, they encountered no such difficulty as those found by Jeffreys and Bernardo; he always got clear-cut decisions from uniform priors without tampering. To see how this was managed, let us examine the simplest of all Laplacian significance tests.

As soon as fairly extensive birth records were kept, it was noticed that there were almost always slightly more boys than girls, the ratio for large samples lying usually in the range $1.04 < (n_b/n_g) < 1.06$. Today we should, presumably, reduce this to some hypothesis about a difference in properties of *X* and *Y* chromosomes (for example, the smaller *Y* chromosome, leading to a boy, would be expected to migrate more rapidly). But for Laplace, knowing nothing of such things, the problem was much simpler. Making no reference to any causal mechanism, he took the model of Bernoulli trials with parameter λ = probability of a boy.

His problem was then: given specific data $D = \{n_b, n_g\}$, do these data indicate the existence of some systematic cause favoring boys? Always direct and straightforward in his thinking, for him the proper question to ask of the theory was simply: Q_L = "Conditional on the data, what is the probability that $\lambda > (1-\lambda)$?" With uniform prior, answer was

$$P_L = \frac{(n+1)!}{n_b!\, n_g!} \int_{\lambda o}^{1} \lambda^{n_b}(1-\lambda)^{n_g} d\lambda$$

with $n = n_b + n_g$, $\lambda_o = 1/2$. In this *Essai Philosophique* Laplace reports many results from this, and in the *Theorie Analytique* (Vol. 2, Chap. 6) he gives the details of his rather tedious methods for numerical evaluation.

Needless to say, Laplace was familiar with the normal approximation to $p(d\lambda|D)$, the inverse of the de Moivre-Laplace limit theorem. But Laplace also realized that the normal approximation is valid only within a few standard deviations of the peak, and when the numbers n_b, n_g become very large, it can lead easily to errors of a factor of 10^{100} in $P_L/(1-P_L)$; hence his tedious methods.

Bernardo's example of Mrs. Stewart's telepathic powers, where the null hypothesis value $\lambda_o = 0.2$ is about 24 standard deviations out, is another instance where the normal approximation leads to enormous numerical errors in *K* (many millions, by my estimate).

But pragmatically, once it is estimated that an odds ratio is about 10^{130}, it hardly matters if the exact value is really only 10^{120} Once it is clear that the evidence is overwhelmingly in favor of H_1, nobody cares precisely how overwhelming it is. After Laplace's time, physicists lost interest in his accurate but tedious evaluations of P_L; for the criterion that we have overwhelming evidence in favor of a positive effect ($\lambda > \lambda_o$), is just that the overwhelmingly greater part of the mass of the posterior distribution $p(d\lambda|D)$ shall lie to the right of λ_o. In the above example, the peak and standard deviation of $p(d\lambda|D)$ are $\lambda' = n_b/n$, $\delta\lambda = [\lambda'(1-\lambda')/n]^{1/2}$ and this criterion reduces to the aforementioned $t = (\lambda' \lambda_o)/\delta\lambda >> 1$, of the modern physicist's significance test-just the same criterion that Jeffreys and Bernardo arrive at in their different ways.

We have noted above that the orthodox t-test and F-test are dominated by Laplace's, and argued that the Jeffreys and Bernardo tests must also be dominated by some other. Let us now compare their specific tests with the ones Laplace would have used in their problems.

5. *Comparisons with Laplace*

In Bernardo's problem we have a normal sampling distribution $p(dx|\lambda,\sigma) \sim N(\lambda,\sigma)$ with σ known. Hypothesis H_o specifies $\lambda = \lambda_o$, H_1 a normal prior $\pi(d\lambda|H_1) \sim N(\mu_1,\sigma_1)$, leading to a normal posterior distribution $p(d\lambda|D,H_1) \sim \mathrm{N}(\lambda',\delta\lambda)$ where

$$(\delta\lambda)^{-2} = n\sigma^{-2} + \sigma_1^{-2} \tag{5.1}$$

$$\lambda' = n(\delta\lambda/\sigma)^2 \bar{x} + (\delta\lambda/\sigma_1)^2 \mu_1 \tag{5.2}$$

Laplace, asking for the probability of a positive effect, would calculate

$$P_L = p(\lambda > \lambda_o|D,H_1) = \Phi(t) \tag{5.3}$$

where $\Phi(t)$ is the cumulative normal distribution, and as always, $t \equiv (\lambda'-\lambda_o)/\delta\lambda$.

Bernardo (Eq. 9) finds for the posterior odds ratio

$$K_B = p(H_o|D)/p(H_1|D) = \exp(-R/2) \tag{5.4}$$

where

$$R = \frac{(\bar{x}-\lambda_o)^2}{\sigma^2/n} - \frac{(\bar{x}-\mu_1)^2}{\sigma_1^2+\sigma^2/n} \tag{5.5}$$

But by algebraic rearrangement, we find this is equal to

$$R = t^2 - w^2 \tag{5.6}$$

where $w \equiv (\mu_1-\lambda_o)/\sigma_1$ is independent of the data and drops out if $\mu_1 = \lambda_o$ or if $\sigma_1 \to \infty$ Bernardo would then find for the posterior probability of the null hypothesis

$$P_B = p(H_o|D) = [\exp(t^2/2) + 1]^{-1} \tag{5.7}$$

and comparing with (5.3) we have, as anticipated, a functional relation $P_B = f(P_L)$. To see the form of it, I plotted P_B against P_L and was surprised to find a quite accurate semicircle, almost as good as one could make with a compass. To all the accuracy one could use in a real problem, the functional relation is simply

$$P_B \cong [P_L(1-P_L)]^{1/2}, \qquad 0 \le P_L \le 1 \tag{5.8}$$

The error in (5.8) vanishes at five points ($0 \le P_L \le 1$).

Since $P_B = f(P_L)$ is single-valued while the inverse function is not, we have the result that Laplace's original significance test does, indeed, dominate Bernardo's. As stressed in Jaynes (1976), one-sided tests always dominate two-sided ones; gives P_L we know everything that Bernardo's K or P_B can tell us; and if $|t| >> 1$ we know in addition whether $\lambda > \lambda_o$ or $\lambda < \lambda_o$, which P_B does not give.

Of course, in this case one can determine that extra bit of information from a glance at the data; so the mere fact of domination is hardly a strong selling point. What is important is that Laplace's method achieves this without any elements of arbitrariness or unBayesianity.

In Jeffreys' problem we have the same sampling distribution, with the standard likelihood function $L(\lambda,\sigma) = \sigma^{-n} \exp[-ns^2Q^2(\lambda)/2\sigma^2]$, where

$$Q(\lambda) \equiv [1 + (\lambda-\bar{x})^2/s^2]^{1/2} \tag{5.9}$$

H_0 and H_1 assign common priors $d\sigma/\sigma$, but H_o specifies $\lambda = \lambda_o$, while H_1 assigns the Cauchy prior $p(d\lambda|\sigma,H_1) = \pi(\lambda|\sigma)d\lambda$ with the density

$$\pi(\lambda|\sigma) = \frac{a\sigma}{\pi(a^2\sigma^2+\lambda^2)} \tag{5.10}$$

scaled on σ (Jeffreys takes $a = 1$, $\lambda_o = 0$, but we define the problem thus to bring out some points noted in Sec. 1). To analyze the import of the data, Jeffreys then calculates the likelihood ratio

$$K_J(\bar{x},s) = \frac{p(D|H_o)}{p(D|H_1)} = M^{-1} \int_0^\infty L(\lambda_o,\sigma)d\sigma/\sigma \tag{5.11}$$

while Laplace (if he used the same prior) would calculate instead the probability of a positive effect, given H_1:

$$P_L(\bar{x},s) = p(\lambda > \lambda_o|D,H_1) = M^{-1} \int_{\lambda 0}^\infty d\lambda \int_0^\infty d\sigma\, \sigma^{-1} \pi(\lambda|\sigma)L(\lambda,\sigma) \tag{5.12}$$

These expressions have a common denominator M, equal to the integral in (5.12) with $\lambda_o = -\infty$.

It is straightforward but lengthy to verify that Jeffreys and Laplace do not ask exactly the same question; i.e., $J \equiv \partial(K_J,P_L)/\partial(\bar{x},s) \neq 0$. However, they are not very different, as we see on making the same approximation (large n) that Jeffreys makes. Doing the σ-integration in (5.12) approximately, the other integrals may be done exactly, leading to the approximate form

$$K_J \cong [\pi(n-1)/2]^{1/2}\, a(1+q^2)/Q^n(\lambda_o) \tag{5.13}$$

where $q \equiv (\bar{x}/as)$. This reduces to Jeffreys' result [Zellner's Eq. (2.7) in this volume] when $a = 1$, $\lambda_o = 0$. In the same approximation, Laplace's result is the tail area of a t-distribution with n-2 degrees of freedom:

$$P_L \cong A_n \int_{\lambda_o}^{\infty} d\lambda/Q^{n-1}(\lambda) \tag{5.14}$$

where A_n is a normalization constant. Of course, if Laplace used a uniform prior for λ, he would find instead the usual "Student" result with (n-1) degrees of freedom.

In the limit of an improper prior ($a \to \infty$), K_J diverges as noted in Sec. 1, the original motivation for both the Jeffreys and Bernardo tamperings; but the arbitrary parameter a cancels out entirely from Laplace's leading term, appearing only in higher terms of relative order n^{-1}.

Had we been estimating λ instead, we should find the result $(\lambda)_{est} = \lambda' \pm \delta\lambda$, where $\lambda' = \bar{x}$, $\delta\lambda = s/\sqrt{n}$. But Laplace's result (5.14) is a function only of the statistic $t = (\lambda'$ $\lambda_o)/\delta\lambda$, and Jeffreys' (5.13) is too for all practical purposes (exactly so if $\lambda_o = 0$, as Jeffreys assumes). Therefore, while considering σ unknown has considerably complicated the mathematics, it does not lead to any real difference in the conclusions. Again, Laplace's test yields the same information as that of Jeffreys, and in addition tells us the sign of $(\lambda-\lambda_o)$. In all cases -Jeffreys, Bernardo, Laplace, and the modern physicist's test- the condition that the data indicate the existence of a real effect is that $|t| >> 1$.

6. *Where does this leave Q_1?*

In summary it should not, in my view, be considered "wrong" to ask the original question Q_1 = "What is the relative status of H_o and H_1 in the light of the data?" But the correct answer to that question depends crucially on the prior range of λ according to H_1; and so the question appears in the retrospect awkward.

Now the original motivation for asking Q_1, stated very explicitly by Jeffreys, was to provide a probabilistic justification for the process of induction in science, whereby sharply defined laws are accepted as universally valid. But as both Jeffreys and Bernardo note, H_o can never attain a positive posterior probability unless it is given some to start with; hence that "pump-priming" lump of prior probability on a single point $\lambda = 0$. It seems usually assumed that this step is the cause of the difficulty.

However, the question Q_1 is awkward in another, and I think more basic, respect. The experiment cannot distinguish differences in λ smaller than its "resolving power" $\delta\lambda = s/\sqrt{n}$. Yet Q_1 asks for a decision between H_o and H_1 even when $|\lambda-\lambda_o| < \delta\lambda$. On the other hand, the experiment is easily capable of telling us whether λ is probably greater or less than λ_o (Laplace's question), but Q_1 does not ask this. In short, Q_1 asks for something which the experiment is fundamentally incapable of giving; and fails to ask for something that the experiment *can* give.

[Incidentally, a "reference prior" based on the Fisher information $i(\lambda)$ is basically a description of this resolving power $\delta\lambda$ of the experiment. That is, the reference prior could be defined equally well as the one which assigns equal probabilities to the "equally distinguishable" subregions of the parameter space, of size $\delta\lambda$. This property is quite distinct from that of being "uninformative", although they happen to coincide in the case of single location and scale parameters].

But what we noted in Sec. 4 above suggests a different view of this. Why does induction need a probabilistic justification if it has already a more compelling pragmatic one? It is for the departures from the previous line of induction (i.e., switching to H_1) that we need -and Laplace gave- a probabilistic justification. Bernardo seems to have sensed this also, in being content with the fact that his $p(H_o|D)$ tends only to 1/2 when H_o is true. Once we see that maintenance of the *status quo* requires no probabilistic justification, the original reason for asking Q_1 disappears.

7. *Conclusion*

What both the Jeffreys and Bernardo tamperings achieved is that they managed to extricate themselves from an awkward start and, in the end, succeeded in extracting the same information from the data (but for the sign of λ-λ_o) that Laplace's question Q_L = "What is the probability that there is a real, positive effect?" elicited much more easily. What, then, was that elusive question Q_2? It was not identical with Q_L, and perhaps does not need to be stated explicitly at all; but in Cox's terminology we may take Q_2 as *any implicate of Laplace's question whose answer is a strict monotonic function of* $|t|$.

We have seen how the answers to seemingly very different questions may in fact convey the same information. Laplace's original test elicits all the information that can be read off from Jeffreys' $K_J(\bar{x},s)$ or Bernardo's $K_B(\bar{x})$. And for all purposes that are useful in real problems, Laplace's P_L may in turn be replaced by the λ' and $\delta\lambda$ of a pure estimation problem. Because of this, I suggest that the distinction between significance testing and estimation is artificial and of doubtful value in statistics-indeed, negative value if it leads to needless duplication of effort in the belief that one is solving two different problems.

D.J. SPIEGELHALTER (*University of Nottingham*):

The papers by Professors Zellner and Siow and Bernardo both suggest reference or 'non-informative' priors for use in Bayes factors, but they produce fundamentally different results. I shall begin by comparing these results, and then discuss the individual merits of the two proposals.

Consider the simple case $\bar{x} \sim N(\mu,\sigma^2/n)$, H_o:$\mu = \mu_o$, H_1:$\mu \neq \mu_o$. Let $\gamma_x = \sqrt{n}(x-\mu_o)/\sigma$ in the notation of Bernardo, who suggests a Bayes factor (17) in favour of the null of $\exp\{-(\gamma_x^2-1)/2\}$, which has behavior similar to that of a significance test. In the case of unknown variance, Zellner proposes the Jeffreys form (2.7) which for n fairly large is approximately equal to $(\pi n/2)^{1/2} \exp(-\gamma_x^2/2)$. For large n, values of γ_x which would lead Bernardo to just reject H_o, would suggest accepting H_o to Zellner.

This is the Lindley paradox, and investigators of Bayes factors have been divided in their support for this phenomenon. Pro-paradox are Zellner (1971), Lindley (1961), Jeffreys (1961), Dickey (e.g. 1971) and Schwarz (1978), while anti-paradox are Akaike (1978), Atkinson (1978), Box and Kanemasu (1973) for 'post data' Bayes factors, and presumably we should include all users of significance tests. Professor Bernardo suggests that a significance test procedure is appropriate in checking a scientific theory. I would be grateful to both authors for some comments on the appropriate practical situations for these two approaches.

The paradox will cause Professor Zellner's Bayes factor wrongly to accept H_o, if the likelihood is concentrated around the true parameter value lying $0(n^{-1/2})$ from H_o. A Bayesian with a true prior under H_1 would, however, consider this event *a priori* extremely unlikely to occur for large n. Moreover, even if this erroneous choice of H_o did occur, for predictive purposes at least, the error is irrelevant since the true model is only a negligible distance from the null. These arguments for the practical use of 'pro-paradox'' Bayes factors are formalised in Smith and Spiegelhalter (1980).

It remains to examine whether the proposals of Zellner and Bernardo are appropriate choices of non-informative prior, within their respective schools of thought on Bayes factors.

Professor Zellner's paper

It has been said at this meeting that 'everything is in Jeffreys'. Perhaps this is an exaggeration, but this paper gives the impression that this work *would* have been in Jeffreys, if only Jeffreys had got round to extending his work to linear models. I trust the authors will take this comment as a compliment of their work, as it is intended.

I have, however, some reservations about the presence of the $\mathbf{X}^T\mathbf{X}$ matrix in the prior specification (3.7b). Changing a prior according to the sampling design would seem somewhat strange. Consider the example of one-way analysis of variance, in which there are I groups with size $n_1, \ldots n_I$, and let $N = \Sigma n_i$. The null hypothesis is of equal group means $H_o: \mu_1 = \ldots = \mu_I = \mu$ against a general alternative. Then (3.12) provides the Bayes factor.

$$K_{01} = \frac{\sqrt{\pi}}{\Gamma(I/2)} \left(\frac{N\text{-}I}{2}\right)^{I\text{-}1/2} (1\text{-}R^2)^{-(N\text{-}I\text{-}1)/2}$$

Consider a prior that does not depend on the sampling design. Considerations of invariance suggest $p(\mu \mid \sigma, H_o) \propto \sigma^{-1}, p(\mu_i \mid \sigma, H_1) \propto \sigma^{-1}$ and $p(\sigma) \propto \sigma^{-1}$ as non-informative priors, which lead to a Bayes factor

$$B_{01} = C_I(\Pi n_i/N)^{1/2}(1\text{-}R^2)^{-N/2}$$

where C_I is some constant of proportionality to be specified. We may adopt the 'device of imaginary results' (Good, 1950) to suggest a plausible value for C_I. Say we observe $R^2 = 0$ (eq ıal group sample means), then we would presumably expect $B_{01} \geq 1$ which implies $C_I^2 \geq N/\Pi n_i$. A lower bound is given when $n_1 = 2, n_i = 1$, $i = 2, \ldots I$, leading to $C_I^2 \geq (I+1)/2$. Assuming this lower bound for illustrative purposes provides a Bayes factor

$$B_{01} = [(I+1)\Pi n_i/2N]^{1/2}(1\text{-}R^2)^{-N/2}$$

To compare the behavior of K_{01} and B_{01}, let $I = 5$, $n_1 = n, n_2 = \ldots = n_5 = m$. Then

$$K_{01} = 1/3\,(N\text{-}5)^2(1\text{-}R^2)^{-(N\text{-}6)/2}$$

$$B_{01} = m^2 3^{1/2}\{n/(n+4m\}^{1/2}(1\text{-}\mathrm{R}^2)^{-N/2}$$

If the design is unbalanced, n being large compared with m, then K_{01} will favour H_o much more than the Bayes factor based on a prior that does not depend on the sampling design. This dependence would appear to be quite important and, as previously mentioned, rather alien to the usual methods of prior specification.

Professor Bernardo's paper

I should first congratulate Professor Bernardo for an ingenious extension of his theory of reference priors to the area of Bayes factors. However, I find the definition (6) of missing information a little forced. If θ has a mixed prior, denoted $p'(\theta)$, should we not seek to maximise $I^{\theta}\{\epsilon, p'(\theta)\}$ with respect to p?

By changing p with n, the author wishes to avoid the situation described by equation (13), in which one accepts H_o as the spread of the prior under the alternative increases. If this prior actually expressed one's beliefs, this behaviour seems quite reasonable. So the objection arises from an inappropriate use of a locally uniform prior, whose ordinate at the likelihood is allowed to go to zero. The problem becomes that of choosing an appropriate ordinate for a locally uniform distribution.

Professor Bernardo's wish to avoid the Lindley paradox would seem appropriate in two contexts at least; when there was a large loss on false rejection of the null, even though the alternative is very close, or when we have strong belief *a priori* in alternatives close to the null. If the latter is true, then this should be modelled in our prior. It can be shown (Smith and Spiegelhalter, 1980) that if the prior shrinks around the null at the same rate as the likelihood concentrates, then one obtains a Bayes factor B_{01} which approximately satisfied

$$-2\log_e B_{01} = \lambda\text{-}(3/2)(p_1\text{-}p_o)$$

where p_i is the number of parameters in H_i, λ is the standard likelihood ratio statistic, and $\lambda \sim \chi^2_{p_1\text{-}p_o}$ under H_o. The multiplier 3/2 compares with the use of 2 by Akaike (1978), 1 by Box and Kanemasu (1973), and $\log_e(n\text{-}p_1)$ by Professor Zellner in expression (3.24).

The example discussed by the author is equivalent to using a multiplier of 1. I am not sure whether the information theoretic argument is to be extended to the general linear model. If so, one should note that the use of a multiplier 1 may lead to a rather strong preference for complex models, since in this case $E[-2 \log_e B_{01}] = 0$ and so the probability that the Bayes factor prefers H_1, given H_o is true, is approximately .5

whatever the complexity of the alternative. I suggest a slightly larger multiplier is more appropriate.

H. AKAIKE (*Institute of Statistical Mathematics, Tokyo*):

We are often told that the Bayesian approach is developed for each particular set of data. This means that the sample size is always equal to 1. I see n's, the sample sizes, in both Professor Zellner's and Professor Bernardo's papers. Any aspects within these papers which essentially depend on n may not then be particularly Bayesian.

In the example of Section 1 of Professor Bernardo's paper, we thus assume $n = 1$. This reduces the problem to the choice of $p = p\{H_0\}$ in relation to the size of σ_1^2, the variance of the prior distribution of the mean μ. For simplicity we assume $\mu_1 = \mu_0 = 0$ and get $p(x|H_0) = N(x|0,\sigma^2)$ and $p(x|H_1) = N(x|0,\ \sigma^2 + \sigma_1^2)$. To keep the predictive distribution $p(x) = pN(x|0,\sigma^2) + (1\text{-}p)N(x|0,\sigma^2 + \sigma_1^2)$ impartial to both $p(x|H_0)$ and $p(x|H_1)$ in terms of entropy, we have to assume $\int p(x|H_0)\log(p(x)/p(x|H_0))dx = \int p(x|H_1)\log(p(x)/p(x|H_1))\,dx$. When $\sigma_1^2 \to \infty$ this will hold only with $p = 0.5$. For this choice of $p = 0.5$, the critical value of x where the posterior probability of H_0 attains 0.5 is almost equal to 2σ for $\sigma_1 = 8\sigma$ and increases to 3σ for $\sigma_1 = 100\sigma$. This seems to suggest that ordinary choice of significance level such as 5% or 1% is fairly reasonable. The fixed choice of the level may further be questioned. But it is now obvious that the ratio σ_1/σ controls the choice.

A.P. DEMPSTER (*Harvard University*):

Both papers are concerned with Bayesian tests of significance. A standard parametric specification depending on parameters (θ,ϕ) is assumed, and the null hypothesis is that θ takes a prespecified "sharp null" value θ_0, while ϕ is unconstrained. Both papers start from the "paradox" of Lindley (1957) who shows that Bayesian testing and tail area testing produce very different judgments when a diffuse prior distribution is assigned to θ given the alternative hypothesis. As sample sizes increase, the diffuse prior implies that the data tend to add much more credence to the null hypothesis relative to standard tail area tests. Both papers develop alternatives to diffuse priors which bring the Bayesian results into relatively close conformity with tail area results. The papers are worthy contributions to theoretical statistics, but in my view irrelevant to statistical practice.

What is probability? Probability does not fall into distinct categories such as subjective, logical, and physical. Any probability model worth using to assess real world uncertainty must command belief, must result from a chain of reasoning, and must not be in clear conflict with known empirical facts. Bernardo appeals to an information-theoretic principle to derive a prior distribution, while Zellner and Siow appeal to plausible postulates originating with Jeffreys. The reasoning behind these derivations is interesting, but there is no way I can commit belief to the resulting prior distributions, since my prior would then depend on the accident of sample size. Also, there is an empirical "how does it work" component to each paper consisting of comparisoı s with tail area results, and suggesting that the disparity between the Bayesian techniques and standard non-Bayesian practice is rather mild. But, since tail area tests are not supposed to be Bayesian, the mildness of the disparity is a logical curiosity rather than evidence that the Bayesian models are credible.

What are significance tests for? The procedures called Bayesian significance testing and tail-area significance testing answer logically different questions, so that the use of the term significance testing for both creates semantic confusion reather than substantive controvery.

In connection with his parable of King Hiero's crown, Savage (1962, pp, 29-33) clearly illustrated the need for Bayesian procedures which provide rational choices between sharp null hypothesis and higher dimensional alternatives. I agree with the Bayesian position which says that the advocates of Neyman-Pearson testing theory are in error when they seek to apply their theory to operational decision-making, as in Pearson (1962). The Neyman-Pearson theory makes probabilistic sense only as a theory *about* tail area tests, and is at that an inadequate theory because if fails to come to grips with the mysteries of conditional testing.

The positive aspect of tail-area tests is that they address real questions which come up in the process of developing a formal model to be used either for purposes of scientific insight or operational decision-making. Specifically, they provide one way to ask whether a nominated model appears to conform to the outside world of fact. Tail-area tests ought to be indispensable to Bayesian statisticians wishing to avoid criticism of their models from two directions. Tail area tests which reject can provide signals that modellers, including Bayesians who have already had their prior model elicited, should go back to the drawing board, because the test shows that the data are trying to say something about phenomena not yet captured in the elicited model. In this case, introspection is not enough, an a further look at the real world may be advisable. Tail area tests which accept can serve to point to possible eventualities whose prior probabilities are influenced only minimally by the data, while these same probabilities may exert a serious influence on later Bayesian conclusions or decisions. In this case, introspection may be all there is, and should be given extra effort. For example, I may not have enough data to detect a significant relation between a chemical agent and a human cancer, but once having raised the question I will not lightly brush off the need to put numbers on prior probabilities of small effects.

In summary, Bayesian statisticians who reject tail area testing are correct when they attack its misuse for decision-making, but are in danger of missing the benefits of correct use in their zeal for things Bayesian. "Significance testing" should be excised from Bayesians have enough good things to do without invading neighboring territory.

What does it mean to "test against an alternative hypothesis"? George Barnard argued at the conference, and I supported him, that significance tests can be valid and important when only the null hypothesis is formulated, as in the Daniel Bernouilli example. Fisher rejected the Neyman-Pearson theory which stressed alternative hypotheses because the theory was couched in terms of long run frequencies, whereas in his mind, as in Daniel Bernouilli's, the purpose of significance testing was to interpret a particular data set. Fisher did not use the formal term alternative hypotheses, but he could scarcely have rejected the concept since the very word "null" suggests that significance testing is a backward way to get at alternative hypotheses.

When a significance test gets to be repeatedly used, appears in "how to do it" books, and becomes distorted by the term "procedure"; then there generally is a reasonably well defined set of alternative hypotheses which are substituted for the null hypothesis when the test produces a significant outcome. It is then sensible, I believe, to use the term "testing a null hypothesis against an alternative hypothesis". I believe also that tail area testing is a clumsy mechanism for the purpose. But I have rejected Bayesian "significance testing" as the answer, so what is left?

The obvious answer in the case of simple null and simple alternative hypotheses is to look ar the likelihood ratio in favor of the alternative. If the ratio is 99 to 1 then the null hypothesis can be "rejected" with similar logic to rejection based on a tail area of .01. When the hypotheses are not simple, my suggestion (1973) is to use the posterior distribution of the likelihood ratio, i.e., the posterior distribution which my Bayesian self would use if I adopted the alternative hypothesis, and to reject the null hypothesis if I am reasonably sure, say 60% sure, that the likelihood ratio is at least 99 to 1. This approach produces judgments similar to tail area tests, and so produces practical answers in the same general range as those of Bernardo and of Zellner and Siow.

These papers resolve the Lindley paradox by producing Bayesian procedures where the paradox largely goes away. I prefer to say there never was a paradox largely because the procedures Lindley contrasts were not comparable in the first place. My work (1973) exhibits alternatives to tail area testing which area genuine significance tests, but are likelihood based. They do not require the contrived priors of Bernardo or Zellner and Siow, but do have a Bayesian element which is relatively insensitive to the choice of prior.

J.M. DICKEY (*University College Wales Aberystwyth*):

Professor Zellner in his paper seems to remain true to Jeffreys' conception when extending Jeffreys' Bayes factors to the general linear model. I should like to point out some disagreeable aspects of the method in Jeffreys' simple context, which extend to the general context. Denote the unknown mean and variance for a simple normal sample, $y_1,\ldots,y_n$, by μ and σ^2. One desires to compare the two models,

$$\mathrm{H}:\mu=0, \qquad \text{versus } H^C:\mu\neq 0$$

As usual, familiar magic words like "knowing little" are used to introduce a particular prior distribution as being worth one's special attention. The idea seems to be to produce an automatic procedure which will be universally accepted. Under H^C, the joint density proposed is

$$p(\mu,\sigma\,|\,H^c) \doteq \{f(\mu/\sigma)/\sigma\}\{K/\sigma\} \tag{1}$$

where

$$f(\mu) = \{\pi(1+\mu^2)\}^{-1}$$

(I have introduced a multiplicative constant K here and written an approximate

equality, relative to the likelihood function, in the sense of Savage's "precise measurement").

I assume that the first bracketed factor in (1) represents the conditional prior information concerning μ given σ,

$$p(\mu\,|\,\sigma,H^c) \doteq \{\pi(1+\mu^2/\sigma^2)\}^{-1}/\sigma \tag{2}$$

(One could argue that my assumption is unwarranted. But an alternative factorization would need to be given, rather than mere magic words). Thus, the second factor would be the marginal prior density for σ under H^c.

$$p(\sigma\,|\,H^c) \doteq K/\sigma \tag{3}$$

My first complaint is that the integrable conditional density (2) is *very special*. I have heard it said that the choice of scale $1.\sigma$ is made "for convenience". But why not 100 σ "for convenience"; or (.001)σ, or (11,682.49) σ? Clearly, the choice should depend on the actual opinion in each application. Should one act against one's opinions and, instead, report a Bayes factor that represents no person's coherent change of opinion?

One may find it difficult thus to specify ones conditional opinion concerning the location conditional on the unknown scale. But what about the *marginal* opinion concerning μ under H^c? Working directly from the joint density (1), we obtain

$$\begin{aligned} p(\mu\,|\,H^c) &= \textstyle\int_0^\infty p(\mu,\sigma\,|\,H^c)d\sigma \\ &\doteq K \textstyle\int_0^\infty f(\mu/\sigma)/\sigma^2\, d\sigma \\ &= K/\,|\mu|. \end{aligned} \tag{4}$$

Again, for my second complaint, this is a very special form and may fail to approximate well ones actual prior opinion concerning μ undr H^c, even locally relative to the likelihood function, even with the constant K open to choice.

Under the hypothesis H, the corresponding prior density which was proposed for σ is

$$p(\sigma\,|\,H) \doteq k/\sigma \tag{5}$$

This constrasts with the conditional density obtained from (1) and (4),

$$\begin{aligned} p(\sigma\,|\,\mu,H^c) &= p(\mu,\sigma\,|\,H^c)\,/\,p(\mu\,|\,H^c) \\ &\doteq |\mu|\,f(\mu/\sigma)/\sigma^2; \end{aligned} \tag{6}$$

which has the asymptotic form near H,

$$\lim_{\mu\to 0}\; p(\sigma\,|\,\mu,H^c) \propto \sigma^{-2} \tag{7}$$

Note, however, that for the new variable $\eta = \mu/\sigma$, η and σ are prior independent under H^c according to (2), and hence for any value of η,

$$p(\sigma|\eta,H^c) \doteq K/\sigma \qquad (8)$$

In particular, (5) and (8) agree for $\eta = 0$, thereby satisfying Savage's condition continuity. (See my discussion to the paper by Professor Smith in these Proceedings). Note that any other point hypothesis, $\mu = \mu_o$, could not be reexpressed in terms of a point hypothesis on η, since $\mu = \mu_o$ means $\eta = \mu_o/\sigma$.

In my paper in press, Dickey (1978), convenient Bayes factors are provided for the normal linear model together with operational methods for use in cases where the likelihood function is more informative than the prior densities. I also treat intersecting hypotheses, as well as nested and unrelated hypotheses.

S. GEISSER(*University of Minnesota*):

In most statistical problems in which one is dealing with a linear regression, the regression arises not from some "true" physical process but largely from a combination of convenience and an adequate fit of the data in hand. The reasons are twofold , first the so-called "true" process governing the data is often very complex and unknown. Secondly, the interest in the data emanates from a need to predict new values rather than to select a "true" physical model. With this view in mind, W. Eddy and I (1979) devised a selection scheme (useful for a variety of situations including linear regression) which is geared to prediction and derives from a Bayes-Non-Bayes methodological compromise. One of its properties, which superficially appears to be unfavorable, is that asymptotically with non-zero probability it can choose a "wrong" higher dimensional model as opposed to a "true" lower dimensional model. However, it turns out that it is approximately equivalent to a Bayesian procedure with penalties (costs or prior weights) that depend on the sample size and the kind of selection error incurred. What this implies is that even if one chooses the higher dimensional model when the lower one is "true", *asymptotically there is no loss incurred* for predictive purposes. I believe that such procedures are more useful for most problems that occur in statistics than those that are geared only to selecting the true model, because of primary interest in prediction and the fact that our net hasn't really been cast over the "true" alternative.

I.J. GOOD (*Virginia Polytechnic and State University*):

In accordance with a theorem of Abraham Wald, a minimax procedure corresponds to a Bayesian procedure with the 'least favorable" prior. I pointed out in Good (1969) that if expected weight of evidence is taken as the utility (or quasi-utility) measure, then Wald's theorem leads to the Jeffreys invariant prior. (I believe this is equivalent to what Dr. Bernardo describes in terms of maximizing the missing information). It gives an explanation of why the reference prior is invariant with respect to mere changes of notation, and also explains why it cannot be entirely satisfactor;: because minimax methods never are entirely satisfactory except possibly against an intelligent opponent. Nature is neither intelligent, nor an opponent, although life is a losing game.

Regarding "Good's paradox", see my contribution to the discussion of Dr. Zellner's paper.

Dr. Bernardo's Table 1, relating tail-area probabilities to Bayes factors, is remarkably consistent with my rough-and-ready rule that a Bayes factor usually lies between 1/(30*P*) and 3/(10*P*) (Good, 1957, p. 863). But, in various applications this formula can be improved; for example, Good and Crook (1974, p. 715), where $N^{1/2}$ comes into the formula.

D.V. LINDLEY (*University College London*):

These two papers bother me. They are extremely thoughtful papers, rich with ideas, yet they fail to adhere to de Finetti's aphorism, "Think about things". If we have a practical problem of data analysis, the quantities have a physical meaning and the scientist knows something about them. He should therefore be encouraged to think about them, or the parameters, and not adopt probability distributions that merely conform to some patterns of ignorance or some formal model. What does he know about θ? Is it really Cauchy? I do not wish to denigrate these papers, for they both help us enormously to understand the way probabilities behave, and are particularly well-written. But, as this conference comes to an end, it does appear to me that we have discussed technicalities too much and that we should balance this necessary activity with some thinking about the *real world, not Greek letters*.

A. O'HAGAN (*University of Warwick*):

Would Professor Bernardo please explain why he chooses the particular limiting process he used in section 2 to obtain equation (11)? If we simply let $\sigma_1^2 \to \infty$ in (9), holding all other quantities fixed, we will obtain posterior odds

$$\frac{\pi(H_1|D)}{\pi(H_0|D)} \to \exp(1/2\,\gamma_x^2)$$

Equation (11) is a consequence of holding γ_1 fixed, so that μ_1 increases with σ_1. In the next section he uses yet another limiting process to reach equation (17). All three limiting processes end with a uniform prior on $(-\infty,\infty)$. All three posterior odds expressions have the same qualitative large-sample behaviour that Professor Bernardo likes. Yet they will give numerically quite different posterior inferences in practice. How are we to choose between them?

A. ZELLNER (*University of Chicago*):

At the 1976 Fontainebleau Conference on Bayesian Methods, I pointed out that Bernardo's procedure for generating prior distributions makes the form of the prior dependent on the likelihood function's form, that is on the design of the experiment. This point, apparently unrecognized by Bernardo, was particularly disturbing to Bernardo and Lindley. In Lindley's discussion of Harold Jeffreys's presentation at the Econometric Society's World Congress meeting in 1970, he termed such a dependence to be incoherent. Jeffreys's, Box and Tiao's and my procedures for generating priors also involve a dependence of a prior's form on the form of the likelihood about which I wrote, Zellner (1977, p. 231) "Since the purpose of a MDIP [maximal data information prior] is to allow the information provided by an experiment to be featured [in the posterior distribution], it seems natural that this form of a MDIP *pdf* that accomplishes this objective be dependent on the design of an experiment". It would be interesting to learn about Bernardo's and Lindley's current position on this issue.

While I do not enjoy raising disturbing points, it should be pointed out that Bernardo's odds ratio in equation (17), $\pi(H_1|D)/\pi(H_0|D) = \exp\{1/2(\gamma_x^2-1)\}$, where $\gamma_x = \sqrt{n}(x-\mu_0)/\sigma$ has a fixed (independent of n) lower bound of $e^{-1/2} = 0.606$. This appears unsatisfactory and is not a characteristic of, for example, Jeffreys's posterior odds ratio for the normal mean problem.

REPLY TO THE DISCUSSION

A. ZELLNER *(University of Chicago)*:

One main objective of Jeffreys's and our work is to provide a coherent framework within which it is possible to rationalize and criticize empirical practice in comparing and choosing between or among hypotheses, for example in the normal mean case, $\lambda = 0$ and $\lambda \neq 0$. In this case Jeffreys (1979) states that "...astronomers had a rough rule that discrepancies up to $\pm 2\sigma$ were likely to disappear with more information, and those beyond χ^2. I was glad to find that these [results] were usually about what my significance tests gave. At least they showed that the rough rule corresponded fairly well to a connected theory". Also, see Jeffreys (1967, p. 273) for another statement of this rough rule, a form of which Jaynes cites approvingly in his comments. Producing a "connected theory" to rationalize sensible rules and to criticize absurd rules for significance testing is one of Jeffreys's and our main objectives which we deem important and intimately related to "real world" significance testing problems, a point which Lindley fails to appreciate in his comments. That significance testing procedures (and other statistical procedures) in physics, astronomy, economics and other sciences are in need of improvement is apparent to many statisticians.

As regards sharp null hypotheses, for which Dempster and Savage, among others see a need and significance tests, Jeffreys (1963) writes, "Every quantitative law in physics implies a series of significance tests that have rejected numerous possible modifications of the law" (p. 409). Similarly in biology, economics and other sciences, significance testing involving sharp null hypotheses plays an important role. Thus, Good's suggestion to "roll together significance testing and estimation into a single process" is misguided in our opinion and contradicts Jeffreys's, Dempster's, Savage's and other's stated "need for Bayesian procedures which provide rational choices between sharp null hyyotheses and higher dimensional alternatives," as Dempster puts it in his comments.

On Jeffreys's and our use of particular Cauchy priors upon which most of our discussants have commented, some of them have apparently missed the point that one of the reasons for their use is that posterior odds ratios based on them rationalize the rough rules used by physicists, astronomers and others in testing. They can represent

prior views in a number of cases and serve as a useful reference prior in others. As Jaynes notes, their use has "the admitted virtue of yielding results that seem reasonable". Thus, in response to Akaike's thoughtful comment, their use leads to Bayesian results which are applicable to many sets of data --a general objective of theorizing in many areas including statistics. Further, as Jeffreys (1967, p. 272) and we stated on the first page of our paper, more informative and/or different priors can, and should be employed if the particular Cauchy priors are deemed inadequate to represent the available prior information. However, we believe that the Cauchy priors which we employed will be found useful in many applications and do serve as a basis to rationalize and criticize much current practice. For example, in our framework p-values are given an interpretation and the implications of a choice of usual critical values for a test statistic can be appraised. In addition, Jeffreys (1967, p. 275) points out that the value of the invariant (divergence) measure,

$$J = \int \log \frac{d\bar{P}}{dP}\, d(\bar{P}-P)$$

for the normal mean problem where P refers to the normal distribution with $\lambda = 0$ and $0 < \sigma < \infty$ and P' to the normal distribution with $\lambda \neq 0$ and $0 < \sigma < \infty$ is $J = \lambda^2/\sigma^2$. He notes that taking a uniform prior on $\theta = \text{arc tan}\,(\lambda/\sigma)$, $-\pi/2 < \theta < \pi/2$ yields exactly the particular Cauchy prior for λ/σ which he employs in the normal mean problem.

We now turn to Jaynes's comments. First, we find no "technical problems" caused by putting a "lump of prior probability on a single point $\lambda = 0$". Second, on the question, "Why should our prior knowledge or ignorance, of λ depend on the question we are asking about it?", Jaynes does not recognize that often when a hypothesis $\lambda = 0$ has been suggested, the value 0 is viewed differently from other possible values. Call this prior information I_0. If the value $\lambda = 0$ is not viewed differently from other values, call this prior information I_1. Then for these frequently encountered circumstances there is good reason for the prior distributions $p_0(\lambda|I_0)$ and $p_i(\lambda|I_1)$ to be different, a fact appreciated by many including Lindley (1965, p. 58 ff.) in his work on testing procedures when prior information is of type I_1.

With respect to the new work of Cox on the logic of questions, we have some doubts about the adequacy of the entropy concept to judge the value of questions. Be that as it may, in our recent work, Zellner and Siow (1979) on the normal mean problem with σ's value unknown, we consider three hypotheses, H_1: $\lambda = 0$. H_2: $\lambda > 0$ and H_3: $\lambda < 0$, with prior probabilities $\pi_1 = 1/2$, $\pi_2 = \pi_3 = 1/4$ and Cauchy priors such as used in our past work, defined over half line $\lambda > 0$ for H_2 and $\lambda < 0$ for H_3. The approximate posterior odds ratios are:

$K_{12} \doteq g(t,\nu)/F(t)$, $K_{13} \doteq g(t,\nu)/F(-t)$ and $K_{23} \doteq F(t)/F(-t)$ where $t = n^{1/2}\,\bar{y}/s$, $g(t,\nu) = (\pi\nu/2)^{1/2}/(1+t^2/\nu)^{\nu-1/2}$, which is Jeffreys's odds ratio given in (2.7) of our paper under discussion and $F(\bullet)$ is the cumulative normal distribution function. It is then the case that the posterior odds ratio for $\lambda = 0$ and $\lambda \neq 0$ (the union of H_2 and H_3)

is just $g(t,\nu)$, Jeffreys's posterior odds ratio. Thus, as Jaynes ingeniously suggested, consideration of two questions $\lambda = 0$ vs. $\lambda > 0$ and $\lambda = 0$ vs. $\lambda < 0$ will yield Jeffreys's result under the *special* prior probabilities given above. However, the practical differences are negligible in this case. Also, the expression for K_{23} above is very close to the result yielded by what Jaynes calls the Laplacian approach. Therefore, when we apply Jeffreys's approach to the three hypotheses, it produces the Laplacian result K_{23}, *as well as* K_{12}, K_{13} and posterior distributions for parameters under all three hypotheses. However, for one-sided alternatives in practice, it is often unreasonable to assume that $\pi_2 = \pi_3$. Our recent work indicates that taking $\pi_1 = .5$, $\pi_2 = .4$ and $\pi_3 = .1$ yields results close to non-Bayesian "one-tailed" testing results in terms of indifference values of t for this normal mean problem when the sample size is about 20.

On predictive distributions and testing, which Jaynes mentions, it is well known that the posterior odds ratio with prior odds ratio equal to one is equal to a ratio of predictive densities and thus a posterior odds ratio of about one indicates close agreement of the predictive densities under the two hypotheses. Also, Jaynes's consideration of values of $|t|$ in appraising hypotheses fails to take adequate account of the role of sample size in evaluating hypotheses. Further, when $|t| << 1$, the important result is that the simpler model (e.g. $\lambda = 0$) can be retained. This is important since it is well known that use of models with redundant or unneeded parameters results in inflation of the mean square error of prediction. Thus, in disagreement with Jaynes, there is obviously something valuable to gain in switching to a simpler model when warranted.

Jaynes remarks that Laplace got "clear-cut decisions from uniform priors". Jeffreys's (1967, p. 128 ff.) discussion of Broad's application of Laplace's rule of succession is relevant. In this case, a uniform prior led to unsatisfactory results in a very basic problem. Jeffreys (1967) comments that, "We really had the simplest possible significance test in our modification of Laplace's theory of sampling, where we found that to get results in accordance with ordinary thought we had to suppose an extra fraction of the initial probability, independent of the size of the class, to be concentrated in the extreme values". (p. 247). See also Geisser's (1978) discussion of this problem. Thus for Jeffreys to get sensible results, it was necessary to use "lumps of probability" on extreme values. Finally, it is surprising to us that Jaynes and Good are apparently in disagreement with Jeffreys and many other scientists and statisticians on the need to distinguish significance testing and estimation.

Spiegelhalter aligns researchers with respect to Lindley's "paradox". This appears to us to be a mistake since there is little paradoxical about Lindley's results. As the sample size increases, good sampling theorists will adjust their significance level in an obvious direction, as pointed out in Zellner (1971, p. 304, fn.) and hence no paradox. Good Bayesians will be familiar with Jeffreys's cogent reasons for and analysis of the dependence of odds ratios on the sample size and again, no paradox. Further, Spiegelhalter requests examples of the use of significance tests in checking scientific theories. The hypothesis of no effect, mentioned in our paper is encountered so frequently that there is no need to publish a list of cases. Also, some theories, for example Milton Friedman's theory of the consumption function predict that parameters will assume particular values and they have been tested extensively in the literature, many times using inadequate testing methodology. For example, there is much confusion about what significance level to employ when the sample size is large, say about 5,000 as in survey data. With such large samples, empirical workers lament that everything looks significantly different from zero at the 5 percent level. Many of them know that they should not be using the 5 percent level but do not know how to adjust it. Some resort to use of p-values which they find hard to interpret. A posterior odds ratio approach provides a clear-cut solution to these problems given that the prior assumptions employed are deemed satisfactory and other subject matter complications are not present —see Jeffreys (1967, pp. 435-436).

With respect to Spiegelhalter's point regarding accepting H_0 when the likelihood is concentrated around the true parameter value lying $0(n^{-1/2})$ from H_0, we agree with him that for large n "the error is irrelevant" and thus question his charge of "to wrongly accept". Also, as many of our discussants and we noted, our prior distributions under alternative hypotheses are informative, not uninformative as stated by Spiegelhalter. They do, however have the property that if the sample evidence violently conflicts with the null hypothesis, posterior distributions for the parameter or parameters under the alternative hypothesis will be very close to what is obtained with a diffuse prior in estimation, a dove-tailing of Jeffreys's testing and estimation results.

On the dependence of our prior on the sample design, this is not unusual. It is also a feature of the Jeffreys, Box-Tiao, Lindley-Bernardo, Zellner and some other priors. Since information in designing an experiment may not be independent of information about parameters's values, such dependence is reasonable. Also, as Box mentioned at this conference session, uninformative and informative are relative terms, relative to the experiment being considered and thus a dependence between prior and design is not unreasonable. In the case of our multivariate Cauchy prior, it can be interpreted as a standard multivariate Cauchy distribution for standardized regression coefficients much like usual beta coefficients. In the case of one independent variable in a regression, the standardized regression coefficient is precisely the unitless quantity $s_x\beta/\sigma$, where s_x is the sample standard deviation of the independent variable, compatible with and a slight generalization of Jefreys's use of λ/σ in the normal mean problem.

In connection with Spiegelhalter's means problem, since the null hypothesis is equality of means, perhaps reflecting prior information that they may not be far different, it is surprising to see that his prior under the alternative has the means uniformly (over the entire real line?) and independently distributed. This prior implies quite strongly that the means may have widely different values and could help to explain Spiegelhalter's problem. In any event, we did not analyze this problem in our paper. For a sensible analysis of the hypothesis of equality of two means with unequal numbers of observations on each, based on Cauchy priors under the alternative hypothesis and with an application to real data, see Jeffreys (1967, p. 278 ff.).

On the issue of the multiplier for p_1 - p_0, as Table 2.1 in our paper referring to the case p_1 - p_0 = 1 shows, the multiplier ln $(n$-1) behaves very reasonably for large n. Also, on choice of models in relation to a loss structure, it is sometimes appropriate to have the loss structure depend on n, as Geisser points out in his comments and this will necessitate a broadened discussion of "the" appropriate multiplier.

In his comments, Geisser describes a frequently encountered circumstance in which investigators are empirically fitting relations with no laws and little or no subject matter theory available. The importance of laws and subject matter theory in science cannot be doubted. But what is one to do in the case described by Geisser? A "starting point" suggested by Jeffreys and others is to consider all variation random until shown otherwise. The hypothesis of "no effect" is thus central as for example in attempting to use a variable to predict stock price changes or gold price changes or in testing a new drug's possible effect. An odds ratio approach seems very appropriate for important problems like these. As regards the Geisser-Eddy predictive scheme, that it provides results that are approximately equivalent to a Bayesian procedure with "penalties (costs or prior weights) that depend on the sample size and the kind of selection error incurred" is very interesting. The afore-mentioned intimate relation of posterior odds ratios and predictive densities, well known to Geisser helps to explain this result. In small samples, however adding too many predictor variables can certainly be harmful in prediction. As the sample size grows, there is a danger that because there is no secure scientific basis for the relationship, it may not be stable. Thus we are back to the desirability of using subject matter theory and laws. On the problem of selecting variables in regression, we have applied the analysis in our paper to the Hald data, also analyzed in the cited Geisser-Eddy paper. We obtained an ordering of models not far different from that of Geisser and Eddy and that based on the residual mean square error criterion. Our results include posterior probabilities for each of the 15 possible models and associated odd ratios. As mentioned at the end of our paper, posterior probabilities have a clear-cut interpretation and can be used to average predictions from alternative models, which may be useful in certain cases and can rationalize *ad hoc* schemes for combining forecasts from alternative models which have appeared in the literature.

With respect to Dickey's remarks, we are at a loss to understand his emphasis on "magic words" and on "automatic procedures which will be universally accepted" in view of our statements regarding prior distributions made on the first page of our paper. Above, we have explained the rationale for the use of our particular Cauchy priors and thus no further comment is needed. Since Jeffreys and we parametrized the normal mean problem in terms of $\eta = \mu/\sigma$ and σ (in Dickey's notation), his equation (8) is relevant and indicates no conlfict between the priors for σ under the null and alternative hypotheses. With respect to other point hypotheses, e.g. $\mu = \mu_0$, at the end of our paper we suggested implicitly that it is possible to write, H_1: $w_i = \epsilon_i$ and H_2: $w_i = \lambda + \epsilon_i$, where $w_i \equiv y_i\text{-}\mu_0$ and to proceed to compute the posterior odds ratio for $\lambda = 0$ vs. $\lambda \neq 0$,using Jeffreys's results without difficulty.

Dempster rejects the use of mechanical tail area testing procedures as we do too. He suggests the use of likelihood ratios. For two simple hypotheses, it is well known that the Bayes factor is equal to the likelihood ratio, while for non-simple hypotheses it is equal to a ratio of averaged likelihood functions. Dempster suggests use of the posterior distribution of the likelihood ratio in testing without providing a clear-cut rationale for his procedure. Is the posterior distribution of the likelihood ratio more fundamentally linked to relative degrees of confidence in competing hypotheses than is the posterior odds ratio? We believe that it is not even though we find the posterior

distribution of the likelihood ratio interesting.

We agree with Good that his "Device of Imaginary Results" is very important. As we noted, Jeffreys used it, without naming it, in the normal mean problem (and many others) to deduce surprising results associated with the possible use of a normal prior for λ. On "Good's Paradox", it is our opinion that it is reflected in Jeffreys's (1967, p. 255) work.

In closing, we thank the discussants for their comments and hope that our responses help to provide a better understanding of the issues which they have raised.

J.M. BERNARDO (*Universidad de Valencia*):

I am most grateful to all discussants for their thought provoking comments. In the following I shall try to answer their queries.

I certainly agree with Professor Jaynes in considering the determination of reference priors a top priority research problem of Bayesian Statistics, and I am obviously flattered that a physicist with a through understanding of statistics finds my result 'a beautifully neat expression with a clear ring of truth to it'. I object however to his description of my derivation as 'chopping away the prior probability of the null until is reduced to what I consider reasonable'. Indeed this is a mathematical consequence of the procedure; but this is obtained from a well defined general theory on reference distributions which has been shown to work in very different situations. I do not need to invent any *ad hoc* procedures, (like Jeffreys-Zellner-Siow do when they arbitrarily choose a Cauchy prior), but I determine the prior which describes the situation in which most remains to be learned from the experiment, and claim that this is a sensible reference point for scientific inference.

This reference prior is *not* a description of the scientist's beliefs, but a description of the situation in which the experiment could conceivably provide more information on the quantity of interest; no wonder that this might depend on the design of the experiment.

Similarly, I do *not* think the procedure consists of a 'mutilation of equations originally designed to answer Q_1, so as to force them to answer instead Q_2'. Indeed, one must specify what it is considered to be the interesting question, i.e., the quantity of interest in my own terminology. If θ were the quantity of interest I would obtain a reference posterior density $\pi(\theta|D)$ for θ. If the question of interest is whether $\theta=\theta_0$ or not. I would obtain a reference posterior probability for $H_0 : \theta=\theta_0$. I dealt with the first question in Bernardo (1979b) and I have tried here to solve the second.

I was very interested in the nearly one-to-one relationship (but for the sign of $\hat{\theta}$) between my reference posterior probability and Laplace's tail area. Indeed, I agree that often the question of interest is whether $\theta>\theta_0$ or not; the corresponding refernce posterior probability is provided in equation (2); see also Bernardo (1979b) in reply to Dawid. However, I do not think that this is the *only* interesting question. I feel it is often convenient in applied work to be able to give a probabilistic description of the plausibility of a sharp null. Confidence levels do *not* have such an interpretation, but reference posterior probabilities do.

Dr. Spiegelhalter wonders what are the appropriate practical situations in which I would use this approach. We all know of those consulting situations in which you are specifically asked to help some people to perform some or other classical test. As a matter of principle, I refuse to do such a thing, but often do not have the time to go on a lengthy full Bayesian analysis. I would then give these people the reference posterior probability of the hypothesis they wanted to test.

About Definition 6, I do not think it is a little forced; for it is a consequence of the fact that, in the present context, the quantity of interest is *not* θ but, say, $\psi=\psi(\theta)$ defined as $\psi=\psi_0$ if $\theta=\theta_0$ and $\psi=\psi_1$ if $\theta\neq\theta_1$ and, thus, we want to maximize the missing information about ψ, *not* that about θ.

I have not yet had time to extend these results to the general linear model. I would very much like however to see the details of Smith & Spiegelhalter method applied to the particular example I discuss. Informal discussion with Professor Smith suggests that both results are numerically very close.

I certainly agree with Professor Geisser that the question of interest is often prediction. If this is the case, one could obtain the appropriate reference predictive distribution: see Bernardo (1979b) in reply to D.J. Bartholomew; no need, I believe, for Bayes-non Bayes compromises. I do not think however that prediction is the *only* possible question of interest. As in the example given by Professor Jaynes, Science often finds it convenient to work in terms of the statistical falsification of new 'simple' working hypothesis.

Professor Dempster finds it difficult to commit belief to a "prior" distribution derived from an information-theoretic principle; we are not arguing however that one should do so. Indeed, we only consider reference priors as technical tools to produce posteriors which are as little affected as possible, in an information-theoretical sense, by prior opinions. On the other hand, we believe that the mildness of the disparity between those Bayesian techniques and some standard non-Bayesian practice is more than a logical curiosity: indeed, some of those classical techniques have been succesfully used in practice, and we would like to understand why, from a *coherent, unified viewpoint*.

Professor Dempster recognizes the need for Bayesian procedures which provide rational choices between sharp nulls and higher dimensional alternatives and its main use as warning signals for modellers; he provides *no* argument however against the use of reference posterior probabilities with such purpose.

It has been said in this Conference that everything is in Jeffreys. Maybe we have to add 'and/or in Good'. Indeed, I am flattered to discover that the numerical outcome of my well-defined procedure is consistent with the rough and ready rule suggested by Professor Good's remarkable intuition.

I do not think it is sensible to assume $n=1$ as Professor Akaike does. By so doing he misses the main point of the discussion, namely the behaviour of the proposed procedures as n increases. One may certainly take $n=1$ if one chooses to call x the vector $x = \{x_1,\ldots,x_n\}$ but then, of course, his argument does not follow. Alternatively one could study the result of using sequentially Akaike's prior: I presume you end up again with Lindley's (or Good's) paradox.

Professor Lindley is certainly right when he mentions the need to think about the real world in order to assess proper prior distributions allowing a subjective Bayesian analysis. I am convinced however that such an analysis is difficult to accept by the scientific community unless it is accompanied by some *reference* result, conditional only to model and data, with which it could be compared. I have tried to provide such a reference for standard problems of hypothesis testing.

Dr. O'Hagan wonders how would one choose among the different limiting processes one can imagine in (9); I think this is bound to depend on the sort of approximation one is interested in. For, (9) is an *exact* expression, which gives the reference posterior probability of the null when $p\,(\mu\,|\,H_1)=N\,(\mu\,|\,\mu_1,\sigma_1^2)$. The status of equation (17) is however very different from that of (11); while (11) is obtained from an *approximation* to the exact expression (9), valid under certain conditions, (17) is another *exact* expression, which gives the reference posterior probability of the null when no distributional assumptions under the alternative are made.

Professor Zellner mentions once more the dependence of the reference prior on the form of the likelihood function, a feature which is common to most approaches to the problem, including his own. I certainly agree with him on the inevitability of this dependence. Professor Lindley's position was recently made explicit in his contribution to the discussion of Bernardo (1979b).

On Professor Zellner's second point, I certainly do *not* regard as disturbing the fact that $\pi(H_0\,|\,D)$ has an upper limit. Indeed, I agree with Professor Jaynes when he questions the need for a probabilistic justification for the maintenance of the *status quo*. The mathematical expression of the fact that, in the absence of evidence against the null, the scientist does not reject H_0, but he is *not* prepared to swear it is true, is the oscillation of $\pi(H_0\,|\,D)$ about 1/2, which we obtain under those conditions. I find this far more reasonable than to expect a convergence to one of $\pi(H_0\,|\,D)$.

REFERENCES IN THE DISCUSSION

AKAIKE, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30**, 9-14.

ATKINSON, A.C. (1978). Posterior probabilities for choosing a regression model. *Biometrika* **65**, 39-48.

COCCONI, G. and SALPETER, E.E. (1958). *Nuovo Cimento* **10**, 646.

COX, R.T. (1946). *Amer. J. Phys.* **14**, 1-13.

— (1961). *The Algebra of Probable Inference.* Baltimore: John Hopkins.

— (1978). Of inference and inquiry. In *The Maximum-Entropy Formalism.* (Levine, R.D. & Tribus, M. eds.) 119-167. Cambridge, Mass.: M.I.T. Press.

DEMPSTER, A.P. (1973). The direct use of likelihood for significance testing. *Proceedings of the Conference on Foundational Questions in Statistical Inference.* (Barndorff-Nielsen, O., Blaesild, P. and Schou, G., eds.) 335-352. University of Aarhus.

DICKEY, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.* **42**, 204-223.

— (1978). Approximate coherence for regression model inference with a new analysis of Fisher's Broadbalk Wheatfield example. *Bayesian Analysis in Econometric and Statistics: Essays in Honor of Harold Jeffreys.* (Zellner, A. ed.) 333-354. Amsterdam: North-Holland.

GEISSER, S. (1980). The contributions of Sir Harold Jeffreys to Bayesian Inference. In

Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys. (Zellner, A., ed.) 13-20. Amsterdam: North Holland.

GEISSER, S. and EDDY, W.F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74**, 153-160.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence.* New York: Haffners.

— (1957). Saddle-point methods for the multinomial distribution. *Ann. Math. Statist.* **28**, 861-880.

— (1969). What is the use of a distribution? In *Multivariate Analysis II.* (Krishaiah, P.R., ed.) 183-203. New York: Academic Press.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/Non Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69**. 711-720.

JAYNES, E.T. (1976). Confidence intervals vs. Bayesian intervals (with discussion). In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* (Harper, W.L. & Hooker, C.A., eds.) 175-257. Dordrecht, Holland: D. Reidel.

— (1979). Marginalization and prior probabilities (with discussion). In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys.* (Zellner, A., ed.) 43-87. Amsterdam: North-Holland.

JEFFREYS, H. (1963). Review of L.J. Savage, et. al. *The Foundations of Statistical Inference,* (1962). *Technometrics* **5**, 407-410.

— (1967). *Theory of Probability.* (3rd rev. ed.). Oxford: University Press.

— (1979). Personal communication.

KLEIN, F. (1939). *The Monist* **39**, 350-364.

LINDLEY, D.V. (1957). A statistical paradox. *Biometrika* **44**, 187-192.

— (1961). The use of prior probability distributions in statistical inference and decision. *Proc. 4th. Berkeley Symposium* **1**, 453-468.

— (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint: Part 2, Inference.* Cambridge: University Press.

MOLINA, E.C. (1963). Some comments on Bayes' Essay. In *Two Papers by Bayes.* (Deming, W.E., ed.) 7-12. New York: Hafner Pub. Co.

PEARSON, E.S. (1962). Some thoughts on statistical inference. *Ann. Math. Statist.* **33**, 394-403.

SAVAGE, L.J (1962). *The Foundations of Statistical Inference: A Discussion.* London: Methuen.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

SHERWIN, C.W., *et al.* (1960). *Phys. Rev. Lett.* **4**, 399-400.

SMITH, A.F.M. and SPIEGELHALTER, D.J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. B.* **42**, 213-220.

WEISSKOPF, V.F. (1961). Selected topics in theoretical physics. In *Lectures in Theoretical Physics,* **3**, 54-105. (Brittin, W., *et al.*, eds.) New York: Interscience Publishers, Inc.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.

— (1977). Maximal data information prior distributions. In *New Developments in the Applications of Bayesian Methods.* (Aykac, A. and Brumat, C., eds.) Ch. 12, 211-232. Amsterdam: North-Holland.

ZELLNER, A. and SIOW, A. (1979). On posterior odds ratios for sharp null hypotheses and one-sided alternatives. *Tech. Rep.* University of Chicago.