

Monographs
on Statistics and
Applied Probability 42

Statistical
Reasoning
with Imprecise
Probabilities

Peter Walley

MONOGRAPHS ON
STATISTICS AND APPLIED PROBABILITY

General Editors

D.R. Cox, D.V. Hinkley, D. Rubin and B.W. Silverman

- 1 Stochastic Population Models in Ecology and Epidemiology
M.S. Bartlett (1960)
- 2 Queues *D.R. Cox and W.L. Smith* (1961)
- 3 Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
- 4 The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
- 5 Population Genetics *W.J. Ewens* (1969)
- 6 Probability, Statistics and Time *M.S. Bartlett* (1975)
- 7 Statistical Inference *S.D. Silvey* (1975)
- 8 The Analysis of Contingency Tables *B.S. Everitt* (1977)
- 9 Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
- 10 Stochastic Abundance Models *S. Engen* (1978)
- 11 Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
- 12 Point Processes *D.R. Cox and V. Isham* (1980)
- 13 Identification of Outliers *D.M. Hawkins* (1980)
- 14 Optimal Design *S.D. Silvey* (1980)
- 15 Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
- 16 Classification *A.D. Gordon* (1981)
- 17 Distribution-free Statistical Methods *J.S. Maritz* (1981)
- 18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
- 19 Applications of Queueing Theory *G.F. Newell* (1982)
- 20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikainen and E. Pesonen* (1984)
- 21 Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)

- 22 An Introduction to Latent Variable Models *B.S. Everitt* (1984)
- 23 Bandit Problems *D.A. Berry and B. Fristedt* (1985)
- 24 Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
- 25 The Statistical Analysis of Compositional Data *J. Aitchison* (1986)
- 26 Density Estimation for Statistical and Data Analysis
B.W. Silverman (1986)
- 27 Regression Analysis with Applications *G.B. Wetherill* (1986)
- 28 Sequential Methods in Statistics, 3rd edition *G.B. Wetherill* (1986)
- 29 Tensor Methods in Statistics *P. McCullagh* (1987)
- 30 Transformation and Weighting in Regression *R.J. Carroll and D. Ruppert* (1988)
- 31 Asymptotic Techniques for Use in Statistics *O.E. Barndorff-Nielsen and D.R. Cox* (1989)
- 32 Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
- 33 Analysis of Infectious Disease Data *N.G. Becker* (1989)
- 34 Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
- 35 Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
- 36 Symmetric Multivariate and Related Distributions *K.-T. Fang, S. Kotz and K. Ng* (1989)
- 37 Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
- 38 Cyclic Designs *J.A. John* (1987)
- 39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
- 40 Subset Selection in Regression *A.J. Miller* (1991)
- 41 Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
- 42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
- (Full details concerning this series are available from the publishers)

Statistical Reasoning with Imprecise Probabilities

PETER WALLEY

*Department of Mathematics
The University of Western Australia
Nedlands, WA 6009
Australia*



CHAPMAN AND HALL
LONDON • NEW YORK • TOKYO • MELBOURNE • MADRAS

UK	Chapman and Hall, 2-6 Boundary Row, London SE1 8HN
USA	Chapman and Hall, 29 West 35th Street, New York NY 10001
JAPAN	Chapman and Hall Japan, Thomson Publishing Japan, Hirakawacho Nemoto Building, 7F, 1-11 Hirakawa-cho, Chiyoda-ku, Tokyo 102
AUSTRALIA	Chapman and Hall Australia, Thomas Nelson Australia, 102 Dodds Street, South Melbourne, Victoria 3205
INDIA	Chapman and Hall India, R. Seshadri, 32 Second Main Road, CIT East, Madras 600 035 First edition 1991

© 1991 Peter Walley

Typeset in India by
Thomson Press (India) Ltd, New Delhi
Printed in Great Britain by
St. Edmundsbury Press, Bury St Edmunds, Suffolk

ISBN 0 412 28660 2 (HB)

All rights reserved. No part of this publication may be reproduced or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, or stored in any retrieval system of any nature, without the written permission of the copyright holder and the publisher, application for which shall be made to the publisher.

British Library Cataloguing in Publication Data

Walley, Peter
Statistical reasoning with imprecise probabilities.
1. Probabilities & statistical mathematics
I. Title
519.2

ISBN 0-412-28660-2

Library of Congress Cataloging-in-Publication Data

Walley, Peter.
Statistical reasoning with imprecise probabilities / Peter Walley.
—1st ed
p. cm.—(Monographs on statistics and applied probability)
Includes bibliographical references and index.
ISBN 0-412-28660-2
1. Mathematical statistics. 2. Probabilities. I. Title.
II. Series.
QA276.W299 1990
519.5—dc20

90-42753
CIP

Contents

Preface	ix
1 Reasoning and behaviour	1
1.1 Introduction	1
1.2 Outline of the book	9
1.3 Interpretations of probability	13
1.4 Beliefs and behaviour	17
1.5 Inference and decision	21
1.6 Reasoning and rationality	26
1.7 Assessment strategies	32
1.8 Survey of related work	42
2 Coherent previsions	52
2.1 Possibility	54
2.2 Probability currency	58
2.3 Upper and lower previsions	61
2.4 Avoiding sure loss	67
2.5 Coherence	72
2.6 Basic properties of coherent previsions	76
2.7 Coherent probabilities	81
2.8 Linear previsions and additive probabilities	86
2.9 Examples of coherent previsions	92
2.10 Interpretations of lower previsions	101
2.11 Objections to behavioural theories of probability	109
3 Extensions, envelopes and decisions	120
3.1 Natural extension	121
3.2 Extension from a field	127
3.3 Lower envelopes of linear previsions	132
3.4 Linear extension	136
3.5 Invariant linear previsions	139
3.6 Compactness and extreme points of $\mathcal{M}(P)$	145

3.7 Desirability and preference	150
3.8 Equivalent models for beliefs	156
3.9 Decision making	160
4 Assessment and elicitation	167
4.1 A general elicitation procedure	168
4.2 Finitely generated models and simplex representations	174
4.3 Steps in the assessment process	177
4.4 Classificatory probability	188
4.5 Comparative probability	191
4.6 Other types of assessment	197
5 The importance of imprecision	207
5.1 Uncertainty, indeterminacy and imprecision	209
5.2 Sources of imprecision	212
5.3 Information from Bernoulli trials	217
5.4 Prior-data conflict	222
5.5 Bayesian noninformative priors	226
5.6 Indecision	235
5.7 Axioms of precision	241
5.8 Practical reasons for precision	248
5.9 Bayesian sensitivity analysis	253
5.10 Second-order probabilities	258
5.11 Fuzzy sets	261
5.12 Maximum entropy	266
5.13 The Dempster-Shafer theory of belief functions	272
6 Conditional previsions	282
6.1 Updated and contingent previsions	284
6.2 Separate coherence	289
6.3 Coherence with unconditional previsions	293
6.4 The generalized Bayes rule	297
6.5 Coherence axioms	301
6.6 Examples of conditional previsions	308
6.7 Extension of conditional and marginal previsions	313
6.8 Conglomerability	317
6.9 Countable additivity	323
6.10 Conditioning on events of probability zero	328
6.11 Updating beliefs	334
7 Coherent statistical models	341
7.1 General concepts of coherence	342
7.2 Sampling models	349

7.3 Coherence of sampling model and posterior previsions	362
7.4 Inferences from improper priors	367
7.5 Confidence intervals and relevant subsets	377
7.6 Proper prior previsions	388
7.7 Standard Bayesian inference	393
7.8 Inferences from imprecise priors	397
7.9 Joint prior previsions	403
8 Statistical reasoning	406
8.1 A general theory of natural extension	408
8.2 Extension to prior previsions	415
8.3 Extension to predictive previsions	421
8.4 Extension to posterior previsions	424
8.5 Posteriors for imprecise sampling models	430
8.6 The likelihood principle	434
9 Structural judgements	442
9.1 Independent events	443
9.2 Independent experiments	448
9.3 Constructing joint previsions from independent marginals	452
9.4 Permutability	457
9.5 Exchangeability	460
9.6 Robust Bernoulli models	467
9.7 Structural judgements	472
Notes	478
Chapter 1	478
Chapter 2	490
Chapter 3	502
Chapter 4	509
Chapter 5	519
Chapter 6	545
Chapter 7	561
Chapter 8	576
Chapter 9	595
Appendices	
A. Verifying coherence	596
B. n -Coherence	599
C. Win-and-place betting on horses	602
D. Topological structure of \mathcal{L} and \mathcal{P}	608
E. Separating hyperplane theorems	611

	CONTENTS
F. Desirability	614
G. Upper and lower variances	617
H. Operational measurement procedures	622
I. The World Cup football experiment	632
J. Regular extension	639
K. W-coherence	642
References	645
Glossary of notation	674
Index of axioms	680
Author index	682
Subject index	688

Preface

When I started writing this book, my mind was full of ignorance and uncertainty, particularly about how best to deal with ignorance and uncertainty. Much of the ignorance and uncertainty remains, now that the book is finished, but it is organized more coherently. I see that as progress.

As the title indicates, the book is about methods of reasoning and statistical inference using imprecise probabilities. The methods are based on a behavioural interpretation of probability and principles of coherence. The idea for such a book originated in 1982, after I had written two long reports on the mathematics and elicitation of upper and lower probabilities. My experience in teaching and applying the existing theories of statistical inference had convinced me that each of them was inadequate. The Bayesian theory is inadequate, despite its great virtues of coherence, because it requires all probability assessments to be precise yet gives little guidance on how to make them. It seemed natural to investigate whether the Bayesian theory could be modified by admitting imprecise probabilities as models for partial ignorance. Is it possible to reconcile imprecision with coherence, vagueness with rationality? Fortunately the answer is yes!

The basic ideas of the book appeared in the two technical reports (1981, 1982). The coherence principles for conditional probabilities and statistical models were worked out in New Zealand, in 1983. A draft of the book was written in 1984–85, under the title ‘Rationality and vagueness’, but at that point I decided that the results were interesting enough to be given a more careful examination. That has taken me four years. The result is that I say much more than I originally intended about other ways of modelling ignorance and uncertainty, and about issues such as finite versus countable additivity, conditioning on continuous variables, the likelihood principle, and the incoherence of standard statistical methods. To keep the book to a manageable length, I have omitted a great deal of material from earlier versions, particularly concerning the aggregation of beliefs, decision making with imprecise utilities, and stronger properties of coherence.

A few comments are necessary on the aims, style and limitations of the book. In writing the book I have tried to work out the implications of a

single, coherent point of view concerning probabilistic reasoning and decision. (In this I have been influenced by the work of de Finetti, although I disagree with him on many important issues.) Perhaps my style suggests greater confidence than I actually feel about controversial issues such as confidence intervals, the nature of aleatory probabilities and the role of countable additivity. The arguments given to support my views are not conclusive, and I am probably wrong about some of these matters. I trust that others will be kind enough to point out my mistakes.

Because the theory of lower and upper previsions is new, I felt it necessary to include a detailed account of the mathematical theory, as well as detailed discussion of interpretations. That makes it difficult to read the book from cover to cover, as I know from my own experience! Readers should first read sections 1.1 and 1.2 for an outline of the main ideas; guidance on where to go from there is given in section 1.2. It may also be profitable to read the summaries at the start of each chapter, before diving into the details.

The mathematical theory in the book is quite general. The theory could have been presented at a more elementary level, by restricting attention to upper and lower probabilities, or to finite spaces and finitely generated models (as in Chapter 4), but in order to build a useful theory of statistical inference it does seem necessary to consider infinite spaces. (The general theory makes use of methods and results from linear functional analysis, whereas the methods of linear programming can be used in the special case of finitely generated models.) The mathematical theory is, I believe, quite simple when one becomes familiar with the basic properties of upper and lower previsions.

The book emphasizes the foundations, principles and rules of probabilistic reasoning, rather than their practical implementation in specific applications. I was tempted to include some more substantial applications of the theory to real problems, as in Walley and Campello de Souza (1986), especially to illustrate how imprecise probabilities can be assessed in practice. However, the complexities of any practical problem demand special assessment strategies and probability models, and an adequate description of these would have added substantially more to an already long manuscript. Readers who are interested in assessment methods and practical implementation could begin by reading Chapter 4, which is concerned with finite models rather than the general theory. Those who want to see specific statistical models should refer to sections 5.3, 5.4, 7.8 and 9.6.

It should be emphasized that much work remains to be done to build a satisfactory theory from the ideas presented here. Most obviously, further work is needed to develop a wider range of imprecise probability models, specific families of robust statistical models, a larger repertoire of assessment strategies to relate probabilities to evidence, a general theory of decision

making when both probabilities and utilities are imprecise, and a theory of group decision. Most importantly, the approach suggested in the book needs to be tested in practical problems.

Many people have contributed to the theory that is developed in the book. My view of probabilistic reasoning has been especially influenced by the writings of Terrence Fine, Bruno de Finetti, Jack Good, J.M. Keynes, Glenn Shafer, Cedric Smith and Peter Williams. On a more personal level, Terrence Fine has been a continual source of wisdom and encouragement, and Luis Raúl Pericchi has managed to impart a little of his energy and enthusiasm.

Substantial portions of the book were written in Nelson, New Zealand (1983), Coventry, England (1984–86), Recife, Brazil (1984–85), Ithaca, USA (1986–87, 1989), Caracas, Venezuela (1988, 1989) and Auckland, New Zealand (1988–89). At different times I was supported by Cornell University (School of Electrical Engineering and Mathematical Sciences Institute), the University of Warwick, Universidade Federal de Pernambuco, Universidad Simón Bolívar, the Department of Scientific and Industrial Research (New Zealand), the British Council, Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil), CONICIT (Venezuela), and the UK Department of Health and Social Security.

I taught one-semester courses based on chapters of the book at Universidade Federal de Pernambuco, Cornell University and Universidad Simón Bolívar, and the book has certainly benefited from this exposure. I am especially grateful to Fernando Campello de Souza, Terrence Fine and Luis Raúl Pericchi for arranging my visits to these universities and for commenting on earlier versions of some chapters, and to Sir David Cox for sending me detailed comments on the penultimate draft of the book. Many of their suggestions have been incorporated in the final version.

I am grateful for the written comments, criticism and encouragement I have received from many other people, including Venkat Anantharam, Susan Andrews, Jim Bergez, Tom Brus, George Casella, Frank Critchley, Gisela García Ramos, Michael Goldstein, David Heath, Peter Huber, Marianela Lentini, Isaac Levi, Dennis Lindley, Isabel Llatas, John Maindonald, Cecilia Loreto Mariz, Steven Morrissey, Robert Nau, José Juis Palacios, Adrian Papamarcou, María Eglée Pérez, Chris Rogers, Bruno Sansó, Alistair Scott, Glenn Shafer, Cedric Smith, María Kathleen Vasconcelos, Larry Wasserman, Peter Whittle, Peter Williams, and two anonymous referees. Jonathan Feuer and Nigel Spencer helped me to analyse the World Cup football data as part of their student projects. Richard Stileman and Elizabeth Johnston of Chapman and Hall gave me a great deal of encouragement during the long process of rewriting. Gail Fish and Val Gladman did a fine job of typing the manuscript.

Margaret Thatcher, Ronald Reagan, Roger Douglas and the monetarist economists deserve some acknowledgement, but no thanks, for showing us that incoherent reasoning can be both persuasive and harmful. Finally, thanks to The Smiths for understanding that these things take time.

Ithaca, New York

CHAPTER 1

Reasoning and behaviour

1.1 Introduction

1.1.1 Probabilistic reasoning

Reasoning begins with the recognition of ignorance and uncertainty. In practical reasoning, whether it is aimed at drawing inferences or making decisions, we need to give appropriate weight both to our uncertainty, about facts or events or consequences of our actions, and to the indeterminacy that arises from our ignorance about these matters. To measure the uncertainty, we need some kind of probability. To measure the indeterminacy, we need imprecise probabilities.

'Imprecise probabilities' will be used as a generic term, to cover mathematical models such as upper and lower probabilities, upper and lower previsions (or expectations), classes of additive probability measures, partial preference orderings, and other qualitative models. Most of the theory in this book is presented in terms of lower previsions, which include the special case of upper and lower probabilities and are roughly equivalent to the other mathematical models.

This book is concerned with probabilistic reasoning, which involves various methods for assessing imprecise probabilities, modifying the assessments where necessary to achieve coherence, updating them to take account of new information, and combining them to calculate other probabilities and to draw conclusions. We are especially concerned with the application of probabilistic reasoning in statistical problems, which we call statistical reasoning, where the goal is to draw conclusions about statistical parameters or future occurrences. To a lesser extent, we are concerned also with decision problems, where the goal is to choose a reasonable course of action; some preferences between actions may be determined by combining imprecise assessments of probabilities and utilities. These topics are of fundamental importance in disciplines as diverse as statistics, probability theory, decision theory, economics, psychology, management science, operations research, philosophy of science and artificial intelligence.

An inevitable consequence of admitting imprecise probabilities is that probabilistic reasoning may produce indeterminate conclusions (we may be unable to determine which of two events or hypotheses is more probable), and decision analysis may produce indecision (we may be unable to determine which of two actions is better). When there is little information on which to base our conclusions, we cannot expect reasoning (no matter how clever or thorough) to reveal a most probable hypothesis or a uniquely reasonable course of action. There are limits to the power of reason.

1.1.2 Fundamental ideas of the theory

The first aim of the book is to construct a mathematical theory of imprecise probabilities, based on a behavioural interpretation and principles of coherence, leading to a theory of conditional probability, statistical inference and decision. In particular, we aim to use the behavioural interpretation to derive the mathematical axioms and rules of imprecise probabilities.

The mathematical theory is based on three fundamental principles: avoiding sure loss, coherence and natural extension. A probability model **avoids sure loss** if it cannot lead to behaviour that is certain to be harmful. This is a basic principle of rationality. **Coherence** is a stronger principle which characterizes a type of self-consistency. Coherent models can be constructed from any set of probability assessments that avoid sure loss, through a mathematical procedure of **natural extension** which effectively calculates the behavioural implications of the assessments.

More general definitions of these three principles, which apply to conditional probabilities as well as unconditional ones, are studied in the second half of the book, and used to construct a theory of statistical reasoning. There is a uniquely coherent rule, called the generalized Bayes' rule, for updating probabilities after observing an event of positive probability. A statistical version of this rule can be used to make inferences about statistical parameters, by constructing imprecise posterior probabilities from imprecise prior probabilities (whose precision reflects the amount of prior information) and a model for how the statistical data were generated (which may also be imprecise, to model physical indeterminacy or to achieve robustness). Other types of statistical reasoning can be modelled through a general concept of natural extension. Frequently used statistical methods involving confidence intervals or 'noninformative' priors often produce incoherent inferences.

1.1.3 Comparison with the Bayesian theory

Our theory has similar goals to the well-known Bayesian theory of statistical inference and decision. In some important respects, we follow the approach

1.1 INTRODUCTION

of de Finetti (1937, 1974). We adopt a behavioural interpretation of probability and prevision, and base our theory on concepts of coherence, albeit different concepts from those of de Finetti. What distinguishes this theory from the Bayesian one is that we admit imprecision in probability and utility models.¹

A central assumption of the Bayesian theory is that uncertainty should always be measured by a single (additive) probability measure, and values should always be measured by a precise utility function. This assumption will be called the **Bayesian dogma of precision**.²

For example, de Finetti assumes that for each event of interest, there is some betting rate that you regard as fair, in the sense that you are willing to accept either side of a bet on the event at that rate. This fair betting rate is your personal probability for the event. More generally, we take your **lower probability** to be the maximum rate at which you are prepared to bet on the event, and your **upper probability** to be the minimum rate at which you are prepared to bet against the event. It is not irrational for you to assess an upper probability that is strictly greater than your lower probability. Indeed, you *ought* to do so when you have little information on which to base your assessments. In that case we say that your beliefs about the event are indeterminate, and that (for you) the event has imprecise probability.

This work is motivated by the ideas that the dogma of precision is mistaken, and that imprecise probabilities are needed in statistical reasoning and decision. The second aim of the book is to explain and defend these ideas. We will examine several formulations, interpretations and defences of the dogma of precision, and argue that none of these is at all convincing. Imprecision, indeterminacy and indecision are compatible with rationality.

1.1.4 Arguments for imprecise probabilities

There are many arguments which support the admission of imprecision. The most important of these can be summarized as follows.

(a) Amount of information

Imprecision in probabilities is needed to reflect the amount of information on which they are based. Suppose, for example, you assess upper and lower probabilities for the event, denoted by A , that a particular thumbtack (drawing pin) will land pin-up on a specific toss. If you have no previous experience with this or similar thumbtacks, the difference between your upper and lower probabilities may be large to reflect your lack of information and consequent caution in betting. If you then observe many tosses of the thumbtack, your upper and lower probabilities will tend to converge as you accumulate information about its propensity to fall pin-up. Probabilities

based on extensive data can be distinguished, through their precision, from those based on ignorance.³

The convergence of upper and lower probabilities as information accumulates can be illustrated by the statistical model described in section 5.3.4. This models an initial state of near-ignorance about the thumbtack, in that your initial lower and upper probabilities for A are zero and one. (Initially you are completely unwilling to bet about the future toss.) Suppose that you observe n tosses in which 20% of the outcomes are pin-up. Your posterior [lower, upper] probabilities for A are then $[0.14, 0.43]$ when $n = 5$, $[0.185, 0.259]$ when $n = 25$, $[0.196, 0.216]$ when $n = 100$, and $[0.1996, 0.2016]$ when $n = 1000$. The imprecision in probabilities is large when the number of observations n is small, but becomes practically negligible when n is sufficiently large.

(b) Complete ignorance

A state of complete ignorance, meaning a total absence of relevant information, can be properly modelled by vacuous probabilities, which are maximally imprecise, but not by any precise probabilities.

(c) Descriptive realism

We know, through introspection, that our beliefs about many matters are indeterminate, and that indecision occurs frequently and is often resolved by arbitrary choice. Imprecise probabilities are needed to model the indeterminacy and indecision.

(d) Elicitation

Imprecise probability models are easier to assess and elicit than precise ones. Any combination of probabilistic judgements will generate an imprecise model. For example, we can construct an imprecise model through qualitative judgements, that some events are probable or that some events are more probable than others, which can be easily understood. The conclusions of probabilistic reasoning can be expressed in a similar way.

(e) Bounded rationality

We may be unable to assess probabilities precisely in practice, even if that is possible in principle, because we lack the time or computational ability, or because we do not know how to analyse a complex body of evidence. We may only be able to evaluate upper and lower bounds for a precise probability. (Or an imprecise assessment may be all that is needed.)

(f) Natural extension

A precise probability model that is defined on some class of events determines only imprecise probabilities for events outside the class.⁴

(g) Conglomerability

If we require that an unconditional probability model can always be coherently extended to any larger domain and to probabilities conditional on any partition, then we must allow the extensions to be imprecise.⁵

(h) Group beliefs and decisions

Groups of individuals do not usually have determinate beliefs or values, even when each individual does, because of disagreement between individuals. So the Bayesian model of precise probabilities and utilities does not apply to groups. In our theory, imprecise probabilities and utilities can be ascribed to both individuals and groups, and both are subject to the same rationality conditions. Disagreement amongst group members over probabilities and utilities can be treated in the same way as conflict between several assessments of one individual: both are sources of imprecision.⁶ More generally, when several sources of information are combined, the extent to which they are consistent can be reflected in the precision of the combined model. (Prior-data conflict (k) is another example of this.)

(i) Statistical inference

The Bayesian approach to statistical inference is to assess precise prior probabilities concerning the statistical parameter, and combine these with the statistical data through Bayes' rule. This results in precise posterior probabilities, even when there is very little prior information and very little information is provided by the statistical data about the parameter. The precision is unwarranted, and this has led many statisticians to reject the Bayesian approach.⁷ An alternative approach is to assess imprecise prior probabilities and combine these with the statistical data through a generalized Bayes' rule. The resulting posterior probabilities are imprecise, and their precision increases with the amount of statistical information.

(j) Robustness

The conclusions of a statistical analysis (or decision analysis) are **robust** when a realistically wide class of probability (and utility) models all lead to essentially the same conclusions.⁸ To check the robustness of an analysis using a precise probability model, one would define a 'realistically wide class' of models by varying those assumptions of the precise model that are arbitrary or unreliable. (For example, independence assumptions might be replaced by approximate independence, Normal distributions by a range of distributions with similar shape.) In effect, this replaces the precise probability model by an imprecise one. Conclusions drawn from the imprecise model are automatically robust, because they do not rely on arbitrary or doubtful assumptions. Of course, these conclusions may be

highly indeterminate, but in that case the conclusions obtained from any precise model will be unreliable. Consider the following example.

(k) *Prior-data conflict*

The central portions of a prior density and statistical likelihood function⁹ can usually be specified more precisely than their tails. The precision of posterior probabilities then depends largely on the extent to which the prior density and likelihood function are consistent. Posterior probabilities will be much less precise, and conclusions much less determinate, when the likelihood function is concentrated in the tail of the prior density. In that case there is said to be **prior-data conflict**.

As an example of this, suppose that a statistical observation x is generated according to a Normal distribution with variance 1 and unknown mean θ . Our prior information indicates that θ is within several units of a specific value μ , and we adopt a prior distribution for θ that is Normal with mean μ and variance 1. The posterior distribution for θ , obtained by applying Bayes' rule, is then Normal with mean $0.5(\mu + x)$ and variance 0.5. Compare the two cases: (i) $\mu = 5.5$, $x = 6.5$, and (ii) $\mu = 3.5$, $x = 8.5$. These yield the same Normal posterior distribution, with mean 6 and variance 0.5. But one would almost certainly want to draw different conclusions in the two cases, and in case (ii) to re-evaluate the weights to be given to the two sources of information, because of the sharp conflict between them. Bayesian conclusions about θ would be robust in case (i) but not in case (ii), with respect to the tails of the prior density. This can be modelled quite easily by adopting imprecise prior probabilities which encompass a range of different tail behaviour. The resulting posterior probabilities would be much less precise in case (ii) than in case (i).¹⁰

1.1.5 Bayesian sensitivity analysis

Bayesians often check the robustness of their inferences and decisions by performing a **sensitivity analysis**. In statistical problems, sensitivity analysis involves assessing realistically large classes of (precise) prior distributions and likelihood functions, combining each pair of functions by Bayes' rule to form a class of posterior distributions, and checking whether these posteriors lead to different conclusions. In decision problems, a class of precise probability models and a class of utility functions are combined in a similar way, to determine a class of optimal decisions.

So sensitivity analysts do admit imprecision in probabilities, and model it by a class of precise probability measures. This is an attractive approach, because of the familiarity and tractability of precise probabilities, and

1.1 INTRODUCTION

because it appears to retain the advantages of the strict Bayesian approach while reducing the arbitrariness in the choice of a single, precise model. But we should be wary of this approach, because the Bayesian justifications for the standard probability axioms are based on the dogma of precision, which is plainly inconsistent with the practice of sensitivity analysis. Apart from familiarity, there is no obvious justification for adopting the Bayesian theory of precise probability as a basis for a theory of imprecise probability.

One way of rationalizing sensitivity analysis is to formulate a weaker version of the dogma of precision, which we call the **dogma of ideal precision**. This asserts that, in any problem, there is an ideal probability model which is precise, but which may not be precisely known. The ideal model might represent your real beliefs, which cannot be precisely elicited because of time limitations or measurement errors. Or the ideal model might be the one that would result from an ideal analysis of the available evidence, if you had the time and intelligence to carry this out. You may be able to ascertain only that the ideal probability model belongs to a specific class of models. Under this interpretation, the imprecision of the adopted model arises only from uncertainty about the ideal model. We will call this the **sensitivity analysis interpretation** of imprecise probabilities.

The dogma of ideal precision is a watered-down version of the dogma of precision. It grudgingly admits imprecision in probabilities, but only because limitations of time and intelligence prevent us from living up to Bayesian ideals. We will argue that the dogma of ideal precision is both unjustified and unnecessary. When there is little relevant evidence, even the 'ideal' probabilities are imprecise.

Consequently, we see no reason to adopt a sensitivity analysis interpretation. We will present the theory in terms of lower and upper previsions, quantities which have a behavioural interpretation as maximum buying prices and minimum selling prices for gambles, rather than in terms of 'ideal' probabilities whose existence and meaning are problematical. The theory is based on principles of coherence that can be justified through the behavioural interpretation and do not rely on the dogma of ideal precision.

Rejecting the dogma of ideal precision opens up the possibility of a theory of statistics and decision that is quite different from Bayesian sensitivity analysis. A third aim of this work, and a continuing theme throughout the book, is to suggest ways in which the behavioural interpretation may lead to departures from Bayesian theory and practice.¹¹

1.1.6 Probability assessment

A final argument for the introduction of imprecise probabilities, as a rational response to scarce or conflicting information rather than an unfortunate

failure to live up to Bayesian ideals, is that it can help us to develop procedures for constructing probabilities from the available evidence.

The relation between probability and evidence has been largely ignored by personalist Bayesians. Personalists admit all additive probability models, even those manifestly inconsistent with the evidence, and give little guidance on how to construct probabilities from evidence. Without such guidance, the choice of precise probabilities is subjective and arbitrary.

A few Bayesians have tried to strengthen the basic axioms by adding further principles that relate probabilities to evidence, aiming to characterize a unique **logical** probability for a hypothesis relative to given evidence. But the resulting logical probabilities are unconvincing, especially those based on complete ignorance. When there is no evidence, any choice of precise probabilities appears to be quite arbitrary.

In this book we will suggest some procedures, which we call **assessment strategies**, that can be used in assessing imprecise probabilities. The idea is to extend the initial domain of interest to include other events or quantities whose probabilities can be assessed more directly, easily, objectively or precisely. These probabilities then have implications for the probabilities of interest, which can be calculated by natural extension. Even if the direct assessments are precise, the probabilities constructed by natural extension will typically be imprecise.

Different assessment strategies can be used in the same problem, and they can be compared according to the precision of the probabilities they produce. Most useful assessment strategies involve conditional probabilities, and many involve statistical models or structural judgements such as independence and permutability. Some simple assessment strategies of these types are studied in the second half of the book.

1.1.7 Summary of aims

In summary, the aims of the book are to:

1. develop a mathematical theory of imprecise probabilities, and its applications to probabilistic reasoning, statistical inference and decision, based on a behavioural interpretation and principles of coherence;
2. compare this with the Bayesian theory, to defend the introduction of imprecision;
3. consider the ways in which a behavioural theory differs from Bayesian sensitivity analysis;
4. suggest some useful models and strategies for assessing imprecise probabilities.

1.2 Outline of the book

In this section we explain how the book is organized. Apart from this introductory chapter, the book can be divided into two halves. Chapters 2 to 5 are concerned with unconditional probability, Chapters 6 to 9 with conditional probability and statistical models. The basic mathematical theory of coherence is presented in Chapter 2 (for unconditional probability), Chapter 6 (for conditional probability) and Chapter 7 (for statistical models). The main ideas concerning philosophy and interpretation are discussed in Chapters 1 and 5, and in sections 2.10, 2.11, 4.1, 6.1, 6.11 and 7.2. This material contains relatively little mathematics, and it should be comprehensible to readers with no mathematical background.

Those interested in the more practical aspects of modelling and working with imprecise probabilities may prefer to start with Chapter 4, and then to read sections 3.9 (on decision making), 6.4 (conditioning), 7.8 (statistical inference), and parts of Chapter 9 (independence and permutability). For examples of imprecise probability models, see especially sections 5.3, 5.4 and 9.6 (on models for Bernoulli trials), 2.9, 4.6, 7.8, and Appendices C and I.

The contents of each chapter are outlined next; fuller summaries can be found at the start of each chapter.

1.2.1 Foundations (Chapter 1)

The aim in this first chapter is to introduce the fundamental ideas of the book (imprecise probabilities, avoiding sure loss, coherence, natural extension and assessment strategies), together with the understanding of probability, reasoning and decision on which these ideas are based. We are especially concerned with defending a behavioural interpretation of probability and identifying principles of reasoning and rationality. We consider probabilistic reasoning in a wider context than in the rest of the book, where the focus is mainly on formalized concepts of coherence.

Section 1.3 presents some basic distinctions between interpretations of probability. The main interpretation developed in this book is characterized as epistemic, behavioural, theoretical, rationalistic and constructive. Section 1.4 outlines a behavioural concept of probability, which is part of a psychological model for explaining or predicting behaviour. Section 1.5 describes the constructive role of probability models in processes of inference and decision.

Sections 1.6 and 1.7 are concerned with norms of internal and external rationality. Internal rationality can be formalized through axioms of avoiding sure loss and coherence. External rationality can be partly

characterized in terms of formal principles, but it also involves the selection and application of assessment strategies. Section 1.8 is an introduction to the literature concerning imprecise probabilities and some alternative models for uncertainty.

1.2.2 Unconditional probability (*Chapters 2–4*)

Chapters 2, 3 and 4 develop a mathematical theory of lower and upper previsions, real-valued functions \underline{P} and \bar{P} that are interpreted as buying and selling prices for gambles. There are no restrictions on the domains of definition of \underline{P} and \bar{P} , but the coherence axioms simplify when the domains are linear spaces. Probability models can be defined on an arbitrary domain, and extended to larger domains by natural extension. It is shown in Chapter 3 that various other models (such as preference orderings on gambles, or classes of linear previsions) are roughly equivalent to upper and lower previsions. Upper and lower probabilities are not sufficiently general as they do not determine upper and lower previsions for all gambles. The elementary properties of coherent upper and lower previsions are derived in Chapter 2.

Chapter 3 develops the mathematical theory in several ways, by introducing the key idea of natural extension, characterizing upper and lower previsions in terms of other mathematical models, and outlining a theory of decision. The theory of unconditional previsions is consistent with Bayesian sensitivity analysis, because there is a one-to-one correspondence between coherent lower previsions and closed convex classes of linear previsions. (Coherent lower previsions are just lower envelopes of such classes.) The mathematical theory in Chapters 2 and 3 is illustrated through simple examples in Chapter 4.

Chapter 4 describes various methods for assessing and eliciting imprecise previsions. Assessment strategies can be built up by combining classificatory or comparative judgements, imprecise assessments of density functions, distribution functions or quantiles, intervals of measures, neighbourhoods of precise probabilities, multivalued mappings, and other types of assessment.

1.2.3 Imprecision (*Chapter 5*)

Chapter 5 discusses some important philosophical issues arising from the introduction of imprecise probabilities. Imprecision can result from incomplete information, as well as from incomplete assessment. We study a statistical model for binary observations, in which imprecision arises from lack of information and from conflict between several types of

information. Bayesian ‘noninformative priors’ are supposed to model an absence of information but cannot do so adequately. We survey the arguments given by Bayesians to support their axioms of precision, and conclude that none of the axioms is justified.

We then examine other theories which incorporate or eliminate the imprecision: Bayesian sensitivity analysis, second-order probabilities, fuzzy sets, Jaynes’ principle of maximum entropy, and the Dempster–Shafer theory of belief functions. On the whole, we judge the present theory to be a more promising foundation for statistical reasoning and decision than any of these alternatives, although sensitivity analysis is closely related to our approach, and belief functions can be regarded as a special type of coherent lower probability.

1.2.4 Conditional probability (*Chapter 6*)

This chapter extends the concept of coherence to conditional previsions, and examines the problem of updating previsions after receiving new evidence. Coherence axioms are derived from two principles, the updating principle and conglomerative principle. When the conditioning event has positive lower probability, conditional previsions are determined by unconditional ones through the generalized Bayes’ rule, a consequence of coherence. This can be regarded as a rule for updating previsions after learning that an event has occurred. (Conditional previsions are completely indeterminate when the conditioning event has probability zero, unless further information is provided.) The requirement that unconditional previsions should be coherent with some conditional previsions, called conglomerability, supports the requirement that precise (finitely additive) probabilities should be countably additive, but it also strengthens the case against precision. There are reasonable lower probability models that are not lower envelopes of countably additive measures.

1.2.5 Statistical reasoning (*Chapters 7–9*)

The last three chapters present a theory of statistical reasoning and inference. The problem considered in Chapter 7 is to characterize coherence of statistical models. A statistical model includes a sampling model, which is a parametrized set of hypotheses about how the statistical data were generated, and may also include prior or posterior previsions, which represent beliefs before or after the data are obtained about the correct hypothesis.

Some popular statistical methods, including Neyman–Pearson confidence intervals and Bayesian inferences from improper priors, produce incoherent

models. On the other hand, a wide range of statistical models are coherent, including Bayesian models based on proper, countably additive priors. Coherent inferences from imprecise priors can be constructed as envelopes of Bayesian inferences, or obtained directly through the generalized Bayes' rule.

Chapter 8 is concerned with statistical reasoning, which is regarded as the application of natural extension to statistical models. In the special case of statistical inference, we extend a sampling model and prior prevision to construct posterior previsions. Because the posterior prevision depends on the sampling model only through the probabilities assigned to the observed data, this procedure satisfies a general statistical principle called the likelihood principle. The other cases examined in Chapter 8 involve natural extension of sampling models and posterior previsions to construct several kinds of prior previsions.

The final chapter is concerned with structural properties of independence, permutability and exchangeability, which play an important role in probability assessment, especially in statistical problems, e.g. in constructing joint from marginal probabilities. The behavioural and sensitivity-analysis approaches lead to different models for judgements of independence and permutability. The differences may be important in practice.

1.2.6 Notes and appendices

Many notes and comments, indicated by superscript numerals such as³, have been added to the body of the text. These notes are used to give references to the extensive literature on the topics we discuss, to refer to discussion in other parts of this book, and to incorporate technical details and qualifications that are not essential to the argument. Readers are advised to disregard the notes on a first reading.

The book has eleven appendices. These discuss topics that are related but peripheral to the main themes of the book. The results in Appendix A are useful in checking whether specified previsions are coherent. Appendix B formalizes coherence conditions that involve only a limited number of gambles. Appendix C applies ideas of coherence to the problem of betting on a horse race. Appendices D and E summarize some mathematical facts, concerning linear topological spaces and separating hyperplane theorems, that are needed in Chapter 3. Appendix F axiomatizes general concepts of desirability and preference. Appendix G defines upper and lower variances and describes their basic properties. Several operational procedures that can be used to elicit imprecise probabilities are compared in Appendix H. One of these procedures was used in an experiment to elicit upper and lower probabilities concerning football games; the results are summarized in

Appendix I. Appendix J describes an alternative method of defining conditional probabilities, and compares it with natural extension. Appendix K compares our concept of coherence with the concept defined by Williams.¹

1.3 Interpretations of probability

The premises and conclusions of most practical reasoning are uncertain. To measure the uncertainty, we need to use some kind of probability. Our aim in this section is to outline our interpretation of probability, by distinguishing it from other possible interpretations.²

We need first to distinguish the issue of interpretation from that of mathematical representation. There are many kinds of mathematical models for uncertainty, such as additive probability measures, upper and lower probabilities, or comparative probability orderings. Any one of these models can be given various interpretations. Similarly, any single interpretation of probability can be given various mathematical representations. In outlining some possible interpretations, we will not presuppose any particular type of mathematical model. The probabilities that are used to express uncertainty in reasoning could be the usual kind of precise probabilities, as in 'the probability of event A is 0.3547', but they could also be imprecise or qualitative, as in 'A is probable' or 'A is more probable than B'.

1.3.1 The importance of interpretation

In order to develop a useful theory of probabilistic reasoning, it seems essential to give a clear, unambiguous interpretation of probability.² Although we have distinguished the issue of interpretation from that of mathematical structure, the two issues are, of course, related. The mathematical structure that is used to represent uncertainty, and the axioms that it is required to satisfy, need to be supported through a specific interpretation.

We will argue here that a particular kind of behavioural interpretation is needed in order for probabilistic conclusions to be useful. In later chapters we will apply this interpretation to justify rationality principles of coherence, and hence to derive a particular mathematical structure for imprecise probabilities. It is difficult to see how this or any other type of probability model could be justified without a clear interpretation of probability. Appeals to tradition or consensus, or to purely mathematical criteria such as simplicity, elegance or tractability, are quite inadequate to show that the usual axioms of probability are appropriate in a theory of reasoning and decision.

A clear interpretation is needed also to guide applications of the theory.

In order to assess the probabilities required in reasoning, and to make sense of probabilistic conclusions, you must understand what these probabilities mean. Otherwise, the assessments and conclusions are arbitrary and apparently useless.

Whereas many mathematicians and probabilists seem to be quite indifferent to issues of interpretation, we regard these issues as essential, both as a starting-point for a mathematical theory and for guiding the later development and application of the theory.

1.3.2 Basic distinctions

Now we outline some basic distinctions between interpretations of probability, in order to classify our own interpretation. The most fundamental distinction is between aleatory and epistemic concepts of probability.³ Aleatory probabilities model randomness in empirical phenomena. Epistemic probabilities model logical or psychological degrees of partial belief, of a person or intentional system. ‘This thumbtack has chance 0.4 of landing pin-up’ is a statement of aleatory probability, because ‘chance’ denotes a physical property that does not depend on the observer. ‘He believes that the thumbtack will probably land pin-up’ is a statement of epistemic probability, because it refers to the beliefs of a specific observer.

Epistemic probabilities (unlike aleatory ones) depend on the available evidence. We can distinguish **logical** interpretations, in which the epistemic probability of a hypothesis relative to a given body of evidence is uniquely determined, **personalist** interpretations, in which probabilities are constrained only by axioms of coherence and not by evidence, and **rationalistic** interpretations, intermediate between the logical and personalist extremes, which require probabilities to be consistent in certain ways with the evidence, without requiring that they be uniquely determined.

Amongst epistemic concepts we can also distinguish **behavioural** interpretations, in which probabilities are interpreted primarily in terms of behaviour (such as betting behaviour, or other choices between actions), from **evidential** interpretations, in which the probability of a hypothesis measures a logical or linguistic relation between the hypothesis and the available evidence.⁴ Personalist interpretations tend to be behavioural, whereas logical interpretations are usually evidential.

The preceding distinction concerns the *meaning* of epistemic probabilities, but it is easily confused with the issue of *measurement*. We can measure (or learn about) probabilities either by **observation** of quantities that they influence, or by **construction** of probabilities from knowledge of the factors that influence them. Thus we can measure epistemic probabilities by

observing choices and assertions (a process of *elicitation*), or by constructing them from the available evidence (a process of *assessment*). It is natural for personalist or behavioural theories of probability to emphasize elicitation as a source of probabilities, and for logical or evidential theories to emphasize assessment, although that is not necessary. (Here we will try to develop a behavioural theory that emphasizes assessment rather than elicitation.) Similarly, we can measure aleatory probabilities either by observing relative frequencies (or other statistical data) that they generate, or by constructing them from measurements of other physical properties to which they are related through known laws.⁵

Finally, we can distinguish operational concepts of probability from theoretical concepts. In **operationalist** interpretations, probabilities are identified with the observations obtained from specified procedures ('meaning' is identified with 'measurement'). Epistemic probability might be identified with a choice of betting rates under well-defined conditions, for example. Aleatory probability might be identified with observed relative frequency.

In **theoretical** interpretations, probabilities model underlying theoretical quantities (beliefs or propensities) that are not directly observable, but which influence observables through their interaction with other theoretical quantities (such as values). Theoretical concepts of probability are usually **dispositional**: probability represents dispositions to behave in certain ways (in the epistemic case), or to produce various outcomes (in the aleatory case).

Again, it is easy to confuse operational with behavioural concepts, but it is important to distinguish them. A behavioural interpretation requires neither personalism nor operationalism.

1.3.3 Epistemic probabilities

This book is mainly concerned with epistemic probability. Some type of epistemic probability is needed to express the uncertainties involved in reasoning and inference. (These probabilities must be epistemic because they depend on the available evidence; reasoning is inherently epistemic.)

What type of epistemic interpretation is needed? The interpretation that we adopt in this book is behavioural, theoretical, rationalistic and constructive.⁶ A behavioural interpretation is needed to tell us how the conclusions of reasoning can be used. (This is argued in sections 1.4 and 1.5.) A theoretical interpretation is needed so that probability models can be used outside the specific context in which they are elicited (sections 1.4 and 2.10). A rationalistic or logical interpretation is needed if the conclusions of reasoning and statistical inference are not to be wholly subjective and arbitrary. Probabilities should be constrained by the available evidence, but

we do not yet have algorithms for determining unique (logical) probabilities from evidence. By developing a rationalistic interpretation, we can move away from a personalist interpretation, towards a logical one. To do so, we must develop assessment strategies for constructing probabilities from evidence (section 1.7). A theory of epistemic probability should therefore emphasize assessment rather than elicitation, even though elicitation (through observed choices) is somewhat easier to integrate into a behavioural theory.

Our interpretation can be compared with the subjective Bayesian interpretation of de Finetti, Savage and Lindley, which is epistemic and behavioural, but also operationalist and personalist, and emphasizes elicitation rather than assessment as a source of probabilities. It can also be compared with the objective Bayesian theories of Jeffreys and Carnap, whose interpretation of probability is epistemic, evidential, theoretical, logical and constructive.

1.3.4 Aleatory probabilities

The probabilities involved in statistical sampling models and scientific theories, which are intended to model randomness in the external world, must be given an aleatory interpretation. Aleatory probabilities appear not only in physical theories such as quantum mechanics and statistical mechanics, but also in biological, economic, sociological, psychological and linguistic theories.⁷ These theories refer to empirical phenomena, not to the beliefs of any person or group of persons.

There are two types of aleatory interpretation, each having several versions. Frequentist interpretations identify probabilities with relative frequencies in finite classes, or with limits of relative frequencies in infinite sequences.⁸ Propensity interpretations take the probability (or chance) of an event to measure its tendency to occur in a particular kind of experiment, and regard probabilities as dispositional properties of experimental arrangements rather than properties of classes or sequences.⁹

Frequentist interpretations are quasi-operationalist and rely on measurement through observation of statistical data. Propensity interpretations are theoretical and allow chances to be constructed from other physical properties through known laws, although typically the required laws are unknown and chances must be inferred from statistical data.

We prefer propensity rather than frequentist interpretations because of their applicability to single (unrepeatable) experiments, and the potential for relating chances to other physical properties. Nevertheless, there are difficult questions for any version of the propensity interpretation: (a) are propensities intrinsic properties of an experiment, or do they depend on our specification of the kind of experiment? (b) are they compatible with

determinism? (c) how are they related to other physical properties? (d) why should chances satisfy the standard probability axioms? and (e) how are they related to epistemic probabilities and rational behaviour? Some answers to these questions will be outlined in section 7.2, although our answers are far from complete.

Aleatory and epistemic probabilities are related through a fundamental rationality principle called the **principle of direct inference**: when you know the values of aleatory probabilities, you should adopt them as your epistemic probabilities and act accordingly.¹⁰ For example, if you know that the chance of a thumbtack landing pin-up is 0.4, then you should be prepared to bet on the outcome ‘pin-up’ at odds of 3 to 2, or at any more favourable odds. Through this principle, all the probability models considered in this book can be given an epistemic, behavioural interpretation. It enables us to discuss statistical inference, whose aim is to learn about unknown aleatory probabilities, entirely in terms of epistemic probabilities.

1.3.5 Summary

The main interpretation of probability that is developed in this book is epistemic, behavioural, theoretical, rationalistic and constructive. This is needed to model uncertainty in reasoning and decision. We also adopt an aleatory interpretation of statistical sampling models. Through the principle of direct inference, aleatory models can always be given an epistemic, behavioural interpretation, so the latter can be applied quite generally.

1.4 Beliefs and behaviour

The mathematical models for reasoning and decision that we consider in this book have two components: probability and utility. We can distinguish three roles for a theory of probability and utility. The probability–utility model can be seen as a psychological model for predicting or explaining the behaviour of humans or other ‘intentional systems’ (discussed in this section), as a constructive strategy for making inferences and decisions (discussed in section 1.5), or as a normative model for evaluating reasoning and decision processes (discussed in sections 1.6 and 1.7).

1.4.1 Intentional systems

It is common in everyday life to suppose that a person’s actions are influenced (and often determined) by their beliefs and values. Beliefs concern matters of fact, states of affairs and future occurrences. In cases of uncertainty, beliefs

are often called partial beliefs, and are mathematically modelled by probabilities.¹ Thus, epistemic probability is a ‘degree of belief’. Values (which are sometimes called desires or objectives) concern states of affairs or consequences of actions, and are modelled by utilities. The reference to beliefs and values (or to probabilities and utilities) is a special kind of psychological explanation for human actions. According to this psychological model, a person prefers one action to another when the first produces greater expected utility, where expected utility is calculated by combining the probability and utility measures.²

We can successfully predict and explain the behaviour of other rational agents, not only humans, in terms of beliefs and values. Dennett (1978, 1987) calls these agents **intentional systems**. The essential properties of intentional systems are that they can make choices or perform actions, they can be regarded as having objectives that their actions are designed to further, and they can be regarded as having some degree of intelligence or rationality (so that their actions are usually chosen appropriately to further their objectives in the light of their beliefs). Intentional systems include groups of persons such as committees, corporations, professional associations and nations, as well as animals, machines, computers, expert systems, or combinations of humans and their computers.³

In the rest of the book, the intentional system to which we ascribe beliefs and values will be called **You**.⁴ This convention is adopted to avoid long-winded references to ‘the intentional system’ or ‘the agent’, and to encourage you (the reader) to consider the theory as a possible model for your own beliefs and behaviour. Most readers are probably human, but the theory could be used to improve the reasoning and decision making of groups, or to design intelligent machines and expert systems.

1.4.2 Beliefs and values as behavioural dispositions

According to the psychological model outlined above, beliefs and values are **behavioural dispositions**: abstract, theoretical states of intentional systems, which can interact in suitable circumstances to produce actions. That is, to have certain beliefs and values is to be disposed to behave in certain ways.⁵

In simple terms, one consequence of Your actions has greater value for You than another when You are disposed to choose the first rather than the second, in cases where you have such a choice. You have a higher degree of belief in one event A than another B when You are disposed to choose a gamble which yields a desirable reward if A occurs, rather than a gamble which yields the same reward if B occurs. By refining these comparisons, and considering more complicated choice behaviour, we can hope to measure

both Your values (in terms of utility models) and Your beliefs (in terms of probability models) by observing Your behaviour.

Two important qualifications are necessary. First, we refer to beliefs as behavioural *dispositions* because they are latent tendencies that can be manifested under suitable conditions, e.g. when You are presented with a choice. We do not identify beliefs with any actual choices or behaviour. Beliefs can be (potentially) manifested in many different ways.⁶ Moreover, beliefs alone do not determine behaviour; they influence behaviour only through their interaction with values.⁷ So beliefs cannot be directly inferred from behaviour, without some assumptions about values. In measuring beliefs, we will try to control values, by establishing a known scale of utility, so that beliefs can be easily inferred from choices. For example, Your probability for an event A could be measured by determining the prices, in units of utility, for which You are disposed to buy or sell a gamble that pays You one unit of utility if A occurs, and nothing otherwise.

The second qualification is that we cannot always explain Your actions in terms of beliefs and values. In many cases Your action may be chosen arbitrarily, and not because it is determined by underlying beliefs and values. (The beliefs and values may themselves be indeterminate.) We must therefore be cautious about inferring the existence of beliefs and values from observed choices; we need to check that the choices reflect real preferences.⁸

1.4.3 Behaviourism

The preceding account of beliefs and values is related to psychological and philosophical theories of behaviourism. It is useful to distinguish several types of behaviourism. **Logical behaviourists**, notably Ryle (1949), identified mental states such as beliefs and values with certain kinds of behavioural dispositions. For Ryle, minds are merely collections of behavioural dispositions.⁹

Logical behaviourism is consistent with the psychological model in section 1.4.2, and with the behavioural interpretation of probability adopted in this book, although it goes somewhat further than we need to. We require only that beliefs and values entail certain behavioural dispositions; there may be more to them than that. For example, some cognitive psychologists have suggested that beliefs have some kind of mental representation. Others have suggested that ‘degree of belief’ is an intensity of feeling or expectation. Beliefs may have effects on thoughts and feelings that are accessible to introspection but not manifested in behaviour.¹⁰

An earlier, more extreme, type of behaviourism was popularized by Watson (1913, 1930) and Skinner (1938, 1953). This is known as **radical behaviourism** or stimulus-response theory. Radical behaviourists were

concerned only with publicly observable behaviour, especially with observable responses to stimuli, and were unwilling to hypothesize unobservable mental states such as beliefs and values to explain this behaviour.¹¹ Moreover, they rejected introspection as a source of evidence about psychological states.

The operationalist theory of probability advocated by de Finetti (1974, 1975), which defines probabilities in terms of responses to stimuli such as scoring rules, is in the radical behaviourist tradition.¹² The basic difficulty with operational definitions of probability is that, if probabilities are identified with specific responses to stimuli, it is not clear why the same probabilities should be given in response to different stimuli, or why they should influence behaviour in other contexts, beyond the specific conditions invoked in their definition. Dispositional concepts of probability and utility are needed to explain consistency in behaviour across a variety of contexts.¹³

1.4.4 *Minimal behavioural interpretation*

The mathematical theory in this book is based on a **minimal behavioural interpretation** of probability: any probability model must have some implications concerning potential behaviour.¹⁴ For example, if You assess the probability of an event to be precisely 0.4, then You should be willing to bet on or against the event at rates arbitrarily close to 0.4 when You are offered the chance. Upper and lower probabilities, or any other measures of ‘degree of belief’ or ‘confidence’, should have a similar behavioural interpretation.

The minimal behavioural interpretation does not require that epistemic probabilities and beliefs can be reduced to behavioural dispositions (they may have the other features mentioned in section 1.4.3), nor that they must be measured through observed behaviour, nor that they must be assessed by thinking about betting or other hypothetical behaviour. It should also be clear that, by adopting the minimal behavioural interpretation, we are not committing ourselves to any version of behaviourism, not even to logical behaviourism. We are requiring only that beliefs and probabilities should (potentially) influence behaviour. That does seem to be an essential part of their meaning, and all the interpretations that have been proposed in previous theories, of both epistemic and aleatory probability, do appear to satisfy the minimal behavioural interpretation.

Aleatory probabilities, for example, do have implications for behaviour through the principle of direct inference, but their meaning obviously goes beyond these behavioural implications – they are primarily models for physical randomness. There may also be more to epistemic probability than can be reflected in behaviour. The mathematical theory of epistemic prob-

ability in this book is based entirely on the minimal behavioural interpretation, but it is sometimes reasonable to extend the minimal interpretation, by adopting (for example) a logical or sensitivity analysis interpretation.¹⁵

1.5 *Inference and decision*

In section 1.4 we outlined the explanatory role of probability–utility models, in psychological explanations for behaviour. They have a second, **constructive** role in reasoning and decision-making processes. That is, strategies for making inferences or decisions involve the assessment of probabilities (and utilities). These strategies will be described in this section. As this book is mainly concerned with constructive methods of reasoning, rather than with psychological explanations of behaviour, it is important to explain why the probability–utility model is likely to be useful in decision making, and why the probabilities used in inference should have a behavioural interpretation.

1.5.1 *Inference versus decision*

Inference is the process of drawing conclusions from premises or evidence. **Decision** is the process of choosing an action. Inference and decision are different kinds of processes. Decision is a more practical and specialized activity, because it has practical consequences and it requires consideration of these consequences and their values. Inference is more theoretical and impersonal.

This distinction is widely recognized, although some have argued that inference should be regarded as a special kind of decision in which You choose between possible conclusions.¹ But one ‘chooses’ a conclusion not because it has greater expected utility than other conclusions, but because it is implied or supported by the evidence. The notions of ‘implication’ and ‘support’ rely on principles of reasoning, not on assessments of utility. It may be that inferences are eventually used as a basis for decision. But at the time they are made, their specific uses may be unknown, and the consequences of choosing a particular conclusion may be wholly indeterminate. (Consider scientific discoveries or even weather forecasts, whose eventual uses are unknown to those who announce them.) It is possible to make valid inferences without considering what use will be made of them.²

1.5.2 *General inference strategy*

In short, our approach to probabilistic inference is to assess probabilities of relevant events, based on all or part of the available evidence, and use

techniques of natural extension to construct ‘posterior’ probabilities concerning the unknown quantities of interest, based on all the evidence. Conclusions are based on these posterior probabilities.

This approach to inference, like the Bayesian approach, involves only probabilities. Utility assessments are not needed. However, the posterior probabilities that result from inference can be used for making decisions, if that is required, by combining them with utility assessments, again by natural extension. One inference, when combined with different utility functions, can be the basis for many different kinds of decision. (Again, consider a probabilistic weather forecast.)

1.5.3 Behavioural interpretation of probability

It is clear that the probabilities involved in inference should be epistemic, because they must depend on the evidence. But why should they be behavioural? The main argument is that, in order to understand the meaning of inferential conclusions, it is important to know how they could be used. But, unless the conclusions have a behavioural interpretation, it is not clear how they can be used for making decisions, or for guiding future inquiry and experimentation.³ An inference should not be confused with a specific decision, but it is important that the inference is potentially useful for purposes of making decisions. If a weather forecaster announces a probability 0.3 of rain tomorrow, then You should be able to combine this number with Your personal utilities to decide whether or not to prepare for a picnic.

As an example of statistical inferences whose behavioural meaning is unclear, consider the Neyman–Pearson theory of confidence intervals. Conclusions about a statistical parameter are expressed as a set of parameter values that can be confidently expected to cover the true value, together with a numerical measure γ called the confidence coefficient. This γ is the aleatory probability of coverage, and hence it is the epistemic probability of coverage before observing the statistical data. But it is not obvious that γ can be interpreted as a posterior probability of coverage, after observing the data. We show in section 7.5 that a behavioural interpretation of γ as a posterior probability is often incoherent with the statistical model. Consequently it is unclear how the Neyman–Pearson inferences should be used. (It seems that they are commonly misinterpreted and misused.)

1.5.4 Evidential interpretations of probability

In the psychological model described in section 1.4, beliefs and values are regarded as pre-existing dispositions, which need to be elicited or inferred

from Your behaviour.⁴ In many cases, however, You do not have pre-existing beliefs and values, or these are too indeterminate to be useful in inference or decision. Instead You need to make some assessments of probabilities by analysing the available evidence. A constructive theory of probabilistic reasoning needs to describe procedures for constructing probabilities from evidence.⁵ For that reason, it may seem that an evidential interpretation of probability should be preferred to a behavioural interpretation.

It is clear that epistemic probabilities are related to both evidence and behaviour. The two main tasks for any theory of epistemic probability are to explain how probabilities should be constructed from evidence, and to explain how probabilities should be used in inference and decision. Which type of theory, behavioural or evidential, is likely to be more successful in accomplishing the two tasks? There are several arguments that favour behavioural theories.

First, the behavioural interpretation of probability can be used to support axioms of coherence, by showing that their violation can lead to irrational behaviour. The axioms entail rules for updating probabilities and making decisions. Behavioural theories can thereby accomplish the second task in a convincing way. It seems also that some kind of behavioural interpretation is necessary if probabilities are to be used in decision making.

Evidential theories, on the other hand, concentrate on the first task, by proposing axioms or conventions to determine probabilities from evidence. That can be done through logical or linguistic relationships between evidence and hypothesis (Carnap), formal representation of ignorance (Jeffreys), maximizing entropy subject to constraints (Jaynes), intuitive recognition (Keynes) or personal judgement (Shafer).⁶ But there is an arbitrary or conventional element in each of these approaches. The authors who have examined even simple types of evidence such as relative frequencies have suggested different rules for translating them into probabilities.⁷ Moreover, the rules are applicable only to simple types of evidence. Evidence can be very complicated, and it seems unlikely that any explicit set of rules could be adequate to cover all types of evidence. So evidential theories seem unlikely to accomplish the first task adequately.

In fact, one can argue that a behavioural interpretation is needed not only to relate probabilities to behaviour, but also to relate them to evidence. It is difficult to see how You can judge whether probability 0.3 or 0.6 is more appropriate to given evidence, without understanding what difference the numbers would make to potential decisions.

These are not really arguments against an evidential interpretation. Rather, they are arguments in favour of a minimal behavioural interpretation. Probabilities should have implications for betting and other behaviour, but they might also be defined by reference to evidence. The evidential

interpretation can be regarded as an addition to the minimal behavioural interpretation, and then it is perfectly consistent with the theory developed in this book.⁸

1.5.5 Decomposition of decision problems

Why should You make decisions by assessing probabilities and utilities? The decomposition of decision problems into separate assessments of probabilities and utilities can be regarded as a fundamental strategy for making decisions. It is not obvious that it is a good strategy.⁹

First we must describe the decision strategy in more detail. We suppose that the decision problem is defined by specifying a set of possible actions. (In practice, this formulation may be an important, and difficult, part of the problem.) The strategy involves constructing a suitable set Ω , whose elements represent possible states of affairs, assessing imprecise probabilities for the states in Ω , and assessing imprecise utilities for the consequences of each possible action under each possible state. The probability and utility models are combined by mathematical rules of natural extension, to construct a partial preference ordering of the actions. When the probabilities and utilities are precise, the actions are ordered according to their expected utility. The case of imprecise probabilities is discussed in section 3.9.

This approach differs from Bayesian decision theory in that generally there will not be an ‘optimal’ action that is preferred to all others. It differs from Bayesian sensitivity analysis in that not all actions that are maximal in the partial ordering need be optimal with respect to some pair of precise probability and utility measures. If there is no optimal action, You may try to sharpen the probability and utility assessments, or repeat the analysis using a different set Ω . Finally, You can choose one of the maximal actions arbitrarily, or by applying some rule such as minimax.

This decision strategy is useful to the extent that it determines preferences between actions, and thereby reduces the set of actions from which You must choose. There is no guarantee that it will be useful, and other strategies may sometimes do better.¹⁰ For example, I have an immediate preference for ordering a strawberry sundae rather than a pineapple one. The introduction of a set Ω , to model my uncertainties about the exact constituents of the sundaes, would complicate the problem and make the decision more difficult. Nevertheless, we can suggest the following reasons why the strategy of assessing probabilities and utilities is often likely to be useful:

1. It often helps to decompose a problem into simpler problems. In this case it may be simpler to think about probabilities and utilities in

isolation, and to use mathematical rules to combine them, than to compare actions directly.

2. It is often quite easy to separate beliefs from values, because we are used to doing so in ordinary discourse.¹¹ The psychological model described in section 1.4.1 is commonly used for explaining and predicting human behaviour.
3. Probabilities may already be available as the result of previous inferences (section 1.5.2), or from outside experts such as weather forecasters.
4. Probability and utility assessments may be useful for other purposes besides making decisions, such as justifying the decision to other people, clarifying its sensitivity to new information, or (in group decision making) separating disagreements over facts from disagreements over values.

We conclude that an epistemic, behavioural concept of probability is needed in inference, and can also be useful in decision, but the construction of preferences from probabilities and utilities is not the only way to make decisions, and it is not always the best way.

1.5.6 Separate theories of probability and utility

Most scientific and statistical work is concerned with inference rather than decision.¹² Inference requires assessments of probability but not utility. It is therefore important to develop a theory of probability that is as separate as possible from considerations of utility and value.¹³ It is not possible to completely eliminate ideas of utility, because we adopt a behavioural interpretation of probability and behaviour always depends on utilities as well as probabilities. But we can restrict attention, in a theory of probability, to cases where utilities are precisely known.

To be more specific, we consider **gambles**, whose possible outcomes are gains or losses of a single commodity. The outcome of a gamble is uncertain and depends on the events of interest. The key assumption is that Your utility for the outcome of a gamble is a linear function of the amount of the commodity that You gain from it. In other words, Your personal utility function for the commodity is precise and linear. This may seem a strong assumption, especially as we have suggested that personal utilities are typically imprecise, but it is realistic for some artificial types of commodity such as lottery tickets. In a lottery with a single valuable prize and many tickets, the value of the tickets You hold can be identified with the chance they give You of winning the prize. If gambles result in gaining or losing lottery tickets, then their utilities can be identified with the number of tickets gained or lost. Thus lottery tickets can serve as a **probability currency**, whose utility for You is precise and linear. (This idea is elaborated in section 2.2.)

It is common practice in science to measure one variable (here, imprecise probability) by constructing artificial situations in which other variables (precise utility) can be controlled. Similarly, we can measure imprecise utilities by controlling probabilities. (You might compare different gambles which yield the consequences of interest with known chances.)¹⁴

The main advantage of separating the theories of probability and utility is that the separate theories are inevitably simpler than a joint theory would be.¹⁵ This is especially important in presenting a theory of statistical reasoning and inference, which relies only on the theory of probability. The main disadvantage is that we must postpone a general theory of decision making, except for the atypical case where utilities can be regarded as precise, which is treated in section 3.9.

1.6 Reasoning and rationality

In section 1.5 we outlined general strategies for inference and decision. There are many ways of implementing these strategies, because there are many ways of assessing probabilities and utilities. So far we have said nothing about which assessments are reasonable. In this section and the next we discuss several kinds of norms which constrain the reasoning process. This leads to a third, **normative** role for a theory of probability and utility, in evaluating the reasoning process and its results.

1.6.1 Reasoning versus rationality

The term ‘reasoning’ emphasizes the constructive aspects of deliberation, whereas ‘rationality’ emphasizes the normative aspects. In very general terms, **reasoning** is a process involving reflective thinking,¹ modelling, analysis, judgement and calculation, aimed at solving problems, reaching conclusions and making decisions. This book is concerned with an important type of reasoning, called **probabilistic reasoning**, which involves assessments of imprecise probabilities.

Rationality is a system of norms and standards for guiding and evaluating reasoning.² We say that the results of reasoning, such as decisions, preferences, conclusions or beliefs, are rational when they are produced by a reasoning process that satisfies these norms. In some cases, when the results are internally inconsistent, we can recognize irrationality from the results themselves, without considering the reasoning process that led to them. For example, intransitive preferences or probabilities that incur sure loss are irrational.

So rationality is a set of standards that characterizes good or valid reasoning. What are these standards? The following, rather general

standards seem to be characteristics of good reasoning. (They apply especially to probabilistic reasoning.)

1. carefully formulating objectives;
2. following methods and strategies that have been useful in analogous problems;
3. using imagination and creativity to reformulate the problem in a more insightful way;
4. decomposing a complex problem into simpler sub-problems;
5. searching for relevant information, and analysing it carefully and dispassionately;
6. formulating and evaluating possibilities (states of affairs, actions, or consequences of actions) with care and imagination;
7. recognizing and modelling both uncertainty and indeterminacy;
8. following reliable strategies, principles and rules where possible, especially to replace intuitive judgements;
9. suspending judgement, rather than making hasty or arbitrary judgements;
10. checking consistency of judgements and models, carefully examining sources of inconsistency;
11. calculating the implications of models, and using these to criticize and revise the models;
12. avoiding biases in judgement and errors in calculation;
13. making the reasoning process as explicit as possible, examining reasons for all steps in the process, considering alternative possibilities and analyses;
14. matching the thoroughness of the analysis to the importance of the problem, simplifying and approximating the unimportant features.

The last recommendation acknowledges that the other criteria, which call for careful, systematic and explicit deliberation, may be more or less appropriate depending on the importance of the problem and the time that can be devoted to it. In practical reasoning there may be insufficient time to live up to the ideals of careful deliberation.

1.6.2 Internal versus external rationality

The recommendations in the preceding list are informal and somewhat vague. To formalize principles of rationality that can be defended as norms, we must consider narrower concepts of rationality. Two types of rationality are considered in this book. The first type, **internal rationality**, is concerned with self-consistency of probability models. Internal rationality will be formalized through principles of avoiding sure loss and coherence. A second concept of **external rationality** is concerned with the conformity of

probability models to the evidence on which they are based.³ In part, external rationality can be reduced to coherence by extending the domain of a probability model to include events which are more closely related to the underlying evidence, but it does involve other principles. External rationality will be discussed in section 1.7.

1.6.3 Avoiding sure loss

Next we describe, informally, our two basic principles of internal rationality: avoiding sure loss and coherence. As a simple paradigm of irrationality, consider intransitive preferences. Suppose that there are commodities X , Y , Z such that You strictly prefer X to Y , Y to Z , and Z to X . (These preferences are said to be **intransitive**.) Strict preference of X to Y can be interpreted behaviourally as a disposition to give up Y , plus something You desire,⁴ in return for X . If You do so, and similarly give up Z in return for Y and X in return for Z , the net result is that You are worse off than initially, because You have lost something You desire. This is known as the **money-pump** argument.⁵

Your behaviour here is irrational because it is harmful to Your own interests. The preferences (behavioural dispositions) that could lead to the behaviour are derivatively irrational. The preferences are not necessarily harmful, but they would be harmful if they were manifested together in behaviour.

In the case where X , Y , Z are gambles whose possible outcomes are expressed in units of precise utility, the preferences express beliefs about the outcomes, and the beliefs can be described as irrational.

This kind of intransitivity in preferences between gambles is a simple example of **incurring sure loss**. The general principle of **avoiding sure loss** is that there should be no finite set of gambles, each of which You are disposed to accept, but whose combination is certain to produce a net loss of utility.

To apply this principle to probabilities, define the **lower probability** of an event A , written as $\underline{P}(A)$, to be the largest price (in units of utility) that You are willing to pay in order to receive 1 unit if A occurs, and nothing otherwise.⁶ Let A^c denote the event that A does not occur. Suppose You assess $\underline{P}(A)$ and $\underline{P}(A^c)$. You are then willing to pay $\underline{P}(A)$ to get 1 if A occurs, and also to pay $\underline{P}(A^c)$ to get 1 if A^c occurs, and the two gambles together give You a net return of $1 - \underline{P}(A) - \underline{P}(A^c)$. To avoid sure loss it is necessary that $\underline{P}(A) + \underline{P}(A^c) \leq 1$, and this is seen to be a fundamental rationality constraint on lower probabilities.⁷

How compelling is the requirement that You should avoid sure loss? Since it is fundamental to our theory, it needs to be carefully examined.

1.6 REASONING AND RATIONALITY

That will be done in Chapter 2, but here we can make some comments of general relevance.

1. The condition applies primarily to behaviour, and derivatively to preferences, beliefs and probability models through their behavioural interpretation. When Your probabilities incur sure loss, it is not certain that You will actually lose, but only that You have dispositions to take actions which would (if carried out) produce a sure loss.
2. To bridge the gap between potential and actual loss, it is often supposed that You are in the position of a bookmaker and there are agents policing Your stated probabilities, who will force You to act on any that incur sure loss, producing a so-called **Dutch book**. That assumption seems unnecessary. It suffices to imagine the gambling scenario as a thought experiment which shows, in an especially vivid way, that 'incurring sure loss' can be harmful, and points to a fundamental inconsistency in Your probability assessments.⁸ (Just as the money-pump argument indicates that intransitivity is a fundamental inconsistency.)
3. A possible objection is that You may be willing to accept any of the gambles in a set individually, but not all together. Gambles will be defined in terms of a special scale of probability currency to meet this objection.

1.6.4 Coherence

Consider again the example of preferences between gambles X , Y , Z . Suppose that You strictly prefer X to Y and Y to Z . The principle of avoiding sure loss implies that You must not prefer Z to X . But it does not imply that You must prefer X to Z . You may have no preference between X and Z . A stronger principle of coherence is needed to establish this preference.

In general, we say that Your beliefs are **incoherent** when there is a finite set of gambles, each of which You are disposed to accept, whose net outcome is certainly no better than the outcome of a further gamble which You are not disposed to accept. The principle of coherence requires You to modify Your beliefs to eliminate this kind of inconsistency. You can do so by dropping some of Your earlier dispositions, or by adding a new disposition to accept the further gamble.

This principle entails transitivity of preferences amongst gambles. Suppose that You prefer X to Y and Y to Z . Then You are willing to accept the gambles $X - Y$ (which is equivalent to giving Y in return for X) and $Y - Z$. Their net outcome is $(X - Y) + (Y - Z) = X - Z$. If Your beliefs are coherent, You are willing to accept the further gamble $X - Z$, which is equivalent to giving Z in return for X . Thus You prefer X to Z .⁹

To illustrate the implications of coherence for lower probabilities, suppose that events A and B are mutually exclusive, and You assess $\underline{P}(A)$, $\underline{P}(B)$ and $\underline{P}(A \cup B)$, where $A \cup B$ denotes the event that either A or B occurs. By the behavioural interpretation of lower probabilities, You are willing to pay $\underline{P}(A)$ to get 1 unit if A occurs, and $\underline{P}(B)$ to get 1 unit if B occurs. The net outcome is equivalent to paying $\underline{P}(A) + \underline{P}(B)$ to get 1 unit if $A \cup B$ occurs. If Your beliefs are coherent, You are willing to do so. Now $\underline{P}(A \cup B)$ is the largest price You are willing to pay to get 1 unit if $A \cup B$ occurs. So we obtain the **super-additivity** constraint, $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$ whenever A and B are incompatible.¹⁰ Similarly, upper probabilities should satisfy the sub-additivity constraint $\bar{P}(A \cup B) \leq \bar{P}(A) + \bar{P}(B)$.

1.6.5 Natural extension

Coherence is a type of self-consistency. A probability model is incoherent if calculating the implications of the model would lead to its modification. That is a much less serious defect than incurring sure loss, and it might arise from analytical limitations rather than fundamental irrationality. For example, when You assess the probabilities of many events, it may be difficult for You to verify whether the assessments are coherent, or to construct a coherent model from them. Fortunately, that can be accomplished through mathematical techniques of **natural extension**. Natural extension is a general method of calculating the behavioural implications of probability assessments. Any assessments that avoid sure loss generate, through natural extension, coherent probabilities for all possible events.

As a simple example of natural extension, suppose that events A and B are incompatible, and You make the assessments $\underline{P}(A) = 0.3$, $\underline{P}(B) = 0.4$. The argument in section 1.6.4 shows that You are effectively willing to pay $\underline{P}(A) + \underline{P}(B) = 0.7$ to get 1 unit if $A \cup B$ occurs. In fact, the value $\underline{P}(A \cup B) = 0.7$ is obtained by natural extension of Your assessments. The natural extension is not the only coherent extension (here, any value of $\underline{P}(A \cup B)$ between 0.7 and 1 would be coherent with the initial assessments), but it is the minimal coherent extension.¹¹ If You had initially assessed $\underline{P}(A \cup B) = 0.5$, without recognizing that this was incoherent with the other assessments, then natural extension would modify this to $\underline{P}(A \cup B) = 0.7$.

Natural extension has a central role in our theory of probabilistic reasoning. It can be used to combine various assessments and components of a probability model into an overall model.

1.6.6 Updating and statistical reasoning

In order to construct a theory of statistical reasoning, it is necessary to extend the concepts of avoiding sure loss, coherence and natural extension

to apply to conditional probabilities and statistical models. The lower probability of event A conditional on event B , denoted by $\underline{P}(A|B)$, can be interpreted either as the updated betting rate on A that You would adopt if You learned only that event B has occurred, or as Your rate for betting on A contingent on B (where the bet is called off unless B occurs). Both interpretations are used in statistical reasoning.

An important task for a theory of probabilistic reasoning is to suggest methods for updating probabilities when new information is received. Provided the new information can be identified with an event, updated probabilities can be constructed by conditioning on the event, through a particular rule of natural extension called the generalized Bayes' rule. Provided the conditioning event has positive lower probability, this rule defines the unique updated probabilities that are coherent with the initial ones.

Techniques of natural extension are especially useful in statistical reasoning. Natural extension can be used to construct some components of a statistical model (posterior or prior or predictive probabilities, or sampling model) from assessments concerning the other components.

1.6.7 The role of mathematical axioms

Much of this book is concerned with the mathematical consequences of the coherence axioms. We believe that it is useful to present the theory through mathematical definitions, axioms and theorems, in order to state both the assumptions and the results clearly and rigorously. However, it is advisable to be wary of any system of axioms, especially of axioms that are proposed as norms of rationality. Such axioms require some compelling justification.¹²

What kind of justification is needed? One might argue that norms of rationality are compelling only to the extent that it is demonstrably in Your interest to obey them. To justify any axiom of rationality, it is necessary to show that violation of the axiom is liable to be harmful to Your interests. This can only be done through a sufficiently clear (behavioural) interpretation of probabilities, or other mathematical concepts to which the axiom refers.¹³

In this book, we attempt to build a theory on rationality axioms that do have a compelling justification, through a behavioural interpretation of probability. That is why so much attention is given, throughout the book, to justification of the coherence axioms. The reader should study these 'justifications' critically!

Avoiding sure loss and coherence are the two basic rationality axioms in our theory of probability. Can stronger axioms be justified? The most obvious candidates are the **axioms of precision** adopted by Bayesians. One

simple version is the additivity axiom: $P(A) + P(A^c) = 1$ for all events A . Assuming that P avoids sure loss, this axiom is equivalent to identity of upper and lower probabilities, or to the axioms of finitely additive probability. Bayesians have put forward various arguments to support these axioms. They will be surveyed in section 5.7. Our conclusion is that the axioms of precision cannot be justified as norms of rationality.

Of course, precise probability models do satisfy our coherence axioms. All finitely additive models satisfy the basic axioms, whereas a stronger axiom of **countable additivity** is required in the standard (Kolmogorov) theory of probability.¹⁴ Should finitely additive models be required to be countably additive? One way to obtain countable additivity is to adopt axioms of **countable coherence** (which strengthen the coherence axioms to apply to some infinite sets of gambles), or **strict coherence** (which rule out sets of marginally desirable gambles that cannot produce a net gain but can produce a net loss). But these stronger axioms are not as compelling as coherence.¹⁵

A stronger argument in favour of countable additivity emerges when we extend the coherence axioms to apply to conditional probabilities. For any model that is finitely but not countably additive, there are problems in which there is no coherent way of defining conditional probabilities, and no way of defining precise conditional probabilities that avoid sure loss. So we believe that finitely additive models which are not countably additive should be ruled out, not on grounds of mathematical convenience (as suggested by Kolmogorov), but because they can lead to unreasonable behaviour.

1.7 Assessment strategies

Under a personalist interpretation, epistemic probabilities are constrained by axioms of coherence, but they are otherwise arbitrary. But some coherent probabilities are plainly irrational because they are inconsistent with evidence. As an extreme example, suppose that You assign probability 0.9 to ‘heads’ (meaning that You are willing to bet on ‘heads’ at odds of 9 to 1 on), on the 1001st toss of an apparently normal coin, after observing only 490 ‘heads’ in the previous 1000 tosses. Such an assessment would seem irrational even if only one of the two sources of evidence (appearance of the coin, or frequency data) was available.

Because of their arbitrariness, personalist probabilities are of little interest in statistical or scientific inference. To reduce or eliminate the arbitrariness, we need principles and guidance for relating probabilities to evidence.¹ In the present rationalistic theory of probability, the links between probabilities and evidence are provided by principles of external rationality and by assessment strategies.

1.7.1 Relevant evidence

The terms ‘evidence’ and ‘information’ refer to the body of available facts, beliefs and judgements that can be used in assessing probabilities. Evidence may include scientific, logical and mathematical knowledge, common-sense beliefs, remembered observations and experiences, the opinions of others, written documents, pictures, mental images and impressions. Items of evidence have various degrees of reliability, ranging from impressions and conjectures to established facts.

Only a relatively small part of this information is likely to be relevant in assessing particular probabilities. At an early stage in the reasoning process, You must judge what part of the evidence is relevant, and what part can be ignored.² We will suppose that You have a fixed body of relevant evidence on which to base Your assessments, but (unlike logical theories) we do not assume that the evidence can be completely specified in a formal language.

1.7.2 Principles of external rationality

One way of relating epistemic probabilities to evidence is through principles of external rationality, to be added to the coherence axioms described in section 1.6. Such principles should include the following.³

(a) Symmetry principles

Symmetries in evidence should be reflected in corresponding symmetries in probabilities. For example, if the relevant evidence concerning a set of experiments is invariant under permutations of the experiments, then so should be the probability model for the experiments. (This property of permutability is discussed in section 9.4.) If the relevant evidence concerning the possible outcomes of an experiment is symmetric in the outcomes, so should be the probability model (see section 5.5.1).

(b) Principle of direct inference

If the evidence includes knowledge of aleatory probabilities then these should be adopted as epistemic probabilities. More generally, if aleatory probabilities are determined by the values of certain parameters, then they should be adopted as epistemic probabilities conditional on the parameter values (see section 7.2.4).

(c) Independence principles

If the evidence includes knowledge that random experiments are physically unrelated, then they should be judged epistemically independent conditional on the values of the parameters that characterize the experiments. For example, if tosses of a thumbtack are known to be physically independent

and identical, with unknown chance θ of landing pin-up, then You should judge the tosses to be epistemically independent conditional on θ (see sections 9.1.2, 9.2.6).

(d) Ignorance principles

When there is little or no relevant evidence, the probability model should be highly imprecise or vacuous. More generally, the precision of probability models should match the amount of information on which they are based (see sections 5.3, 5.5).

These principles are useful, but in many problems they are not directly applicable.⁴ In statistical problems, the principle of direct inference can be used to determine epistemic probabilities conditional on each statistical hypothesis, but gives no direct help in assessing the prior probabilities of the hypotheses. In such cases You might have relevant prior information, but You need some way of translating it into probabilities.

1.7.3 Inductive logic

Theories of inductive logic have aimed to completely formalize probabilistic reasoning, by adding further axioms to characterize a unique **logical** probability for any hypothesis on given evidence. The most fully developed theory of this kind is due to Carnap (1952, 1962, 1971, 1980).⁵ Carnap (1952) augmented the basic axioms of coherence and precision by adding various symmetry and invariance axioms, and principles that regulate learning from relative frequencies. In simple problems concerning hypotheses and evidence that can be specified in an elementary language, the whole set of axioms determines a system of logical probabilities that is unique up to a single real parameter.

A major defect of Carnap's theory is that probabilities depend not only on the empirical meaning of hypothesis and evidence, but also on the formal structure of the language in which they are expressed.⁶ For example, let A_j denote the event that the j th object examined from a sequence of objects is red. Then $P(A_1)$ and $P(A_2|A_1)$ depend on whether the predicates of the language are taken to be {red, non-red} or {red, blue, green, ...}.

This defect seems unavoidable in any theory which assumes that all probabilities, even those based on ignorance, are precise. Indeed, Carnap's work shows that, if it is assumed that logical probabilities are precise, they cannot have all the invariance properties that one would expect them to have.⁷ These are further indications that the axiom of precision is unjustified.

There are three important theories of logical probability, due to Keynes (1921), Kyburg (1974a) and Levi (1980), which do allow logical probabilities

1.7 ASSESSMENT STRATEGIES

to be imprecise. The difficulty is that, without the axiom of precision, principles of external rationality such as those in section 1.7.2 are not nearly strong enough in general to determine unique (imprecise) probabilities.⁸

In any case, logical theories of probability are useful only to the extent that they can provide rules or guidance that enable us, in practical problems, to construct the logical probabilities. It is of little use to know that unique logical probabilities 'exist' if they cannot be calculated.⁹ As the axioms of inductive logic and the principles in section 1.7.2 seem inadequate to determine probabilities based on realistically complex bodies of evidence, we need to provide other, constructive methods for relating probabilities to evidence.

1.7.4 Assessment strategies

The motivation for assessment strategies is to replace difficult or intuitive judgements of probabilities, as far as possible, by judgements that are more firmly grounded in evidence. That can often be done by building a probabilistic or statistical model, and using this to calculate upper and lower bounds for the probabilities of interest.¹⁰

As a well-known example, suppose that You are at a party with 24 strangers, and (having nothing better to do) You wish to assess an epistemic probability that there are at least two people at the party who have the same birthday. One assessment strategy is to make an intuitive assessment (or 'guess') of the probability. An alternative is to make some simple assumptions about the distribution of birthdays over the year and the independence of birthdays amongst people at the party, and use this model to calculate the probability of interest. The second strategy is more reliable than the first. (In fact, intuitive assessments of this probability tend to be very much lower than those calculated from a reasonable model.)

This kind of model building is quite familiar to statisticians and others who engage in probabilistic reasoning. It suggests a general way of thinking about the reasoning process, as the selection and application of **assessment strategies**. These can be broken down into four steps:

1. identify some events or unknown quantities, called **targets**, whose probabilities need to be assessed;
2. select other events or quantities, called **carriers**, which are related to the targets, but related more directly than the targets to the evidence;
3. use the evidence (and principles of external rationality) to make imprecise probability judgements concerning the carriers;
4. calculate imprecise probabilities concerning the targets from those concerning the carriers, by natural extension.

The choice of a particular strategy, step 2, is made in the light of both 3 and 4. One aim is to determine the target probabilities as precisely as possible. A second aim is to reduce the need for intuitive or arbitrary judgements.¹¹ The carriers are chosen because their probabilities can be assessed more reliably or precisely, through objective judgements, statistical models and the principles of external rationality. Some intuitive judgements may still be needed in steps 2 and 3 of this process, but step 4 can be carried out by applying formal rules of natural extension.

1.7.5 Thumbtack example

To illustrate the use of assessment strategies in a simple example, suppose that the target event, denoted by A , is that a particular thumbtack lands pin-up on its next toss. The relevant evidence would typically consist of records of previous tosses of this thumbtack, records (or vague recollections) concerning other thumbtacks, and observation of its physical structure to suggest both its bias and its similarity to other thumbtacks. We can suggest several assessment strategies for using this evidence to construct upper and lower probabilities of A , denoted by $\bar{P}(A)$ and $\underline{P}(A)$.

1. Make direct, intuitive assessments of $\bar{P}(A)$ and $\underline{P}(A)$. For example, You might choose these to be upper and lower bounds for the observed relative frequency of pin-ups in previous tosses, r , and choose the degree of imprecision $\bar{P}(A) - \underline{P}(A)$ to be a decreasing function of the number of previous tosses, n . This strategy might be adequate for large n , when both $\bar{P}(A)$ and $\underline{P}(A)$ are approximately equal to r , but for small n it would be difficult to make intuitive assessments, and a more detailed model would be needed.
2. You could separate the evidence about the n tosses of this thumbtack from the other ('prior') evidence as follows. Using only the prior evidence, assess imprecise probabilities concerning the outcomes of $n + 1$ tosses. It would be reasonable, using the symmetry principles, to regard the n tosses, using the generalized Bayes' rule. If the initial model was $P(A)$ by conditioning this model on the observed outcomes of the first n tosses, using the generalized Bayes' rule. If the initial model was exchangeable, this strategy would be mathematically equivalent to the next one, although the next one may be easier to apply.¹²
3. Based on Your experience with thumbtacks, You might model successive tosses as physically independent trials with some constant chance θ of landing pin-up. The epistemic probability of the observed outcomes conditional on θ is then determined through the independence principle and principle of direct inference. You could also assess prior probabilities

concerning θ . If there was little other evidence, the ignorance principle could be invoked to support a type of near-ignorance prior. Natural extension of the statistical model and prior probabilities then determines $\bar{P}(A)$ and $\underline{P}(A)$ as posterior probabilities. For large n , $\bar{P}(A)$ and $\underline{P}(A)$ will nearly agree, even for a near-ignorance prior, but for small n they may be quite different. (See sections 1.1.4 and 5.3.4 for numerical examples.)

4. You could extend the statistical model to make fuller use of the evidence about other thumbtacks. One way to do so is to assume that different thumbtacks have different chances θ which are drawn independently from a distribution parametrized by ϕ . Evidence about the physical similarities between thumbtacks would be used to assess (highly imprecise) prior probabilities concerning ϕ . These would be combined with the statistical model, by natural extension, to produce upper and lower probabilities for A conditional on the outcomes of tossing all the thumbtacks.¹³

This example is more straightforward than most of the assessment problems that confront us, but it does illustrate some general features of assessment. First, we will try to apply the principles of external rationality (section 1.7.2) whenever possible. If some of the evidence can be treated as statistical data, then we will try to construct statistical models for how the data were generated, and apply the independence principle and principle of direct inference. Strategy 3, which does this, seems preferable to strategy 2, which does not. However, this strategy is justified only to the extent that the statistical model is reliable. In order to reliably specify the form of the distribution for θ in model 4, You would need many observations from many different thumbtacks. If these were available, then strategy 4 would be preferable to 3 because it makes fuller use of the evidence and would tend to produce greater precision in probabilities for A . If not, imprecise or 'robust' models could be used in 4, and the comparison between 3 and 4 would depend on the extent of this imprecision. Finally, when there is little prior information, we will adopt highly imprecise prior probabilities, such as the near-ignorance models described in section 5.3.

1.7.6 Types of assessment strategy

Many types of assessment strategies will be described in this book. These include methods for the following:

1. constructing upper and lower previsions from assessments of upper and lower probabilities (described in section 3.2);
2. constructing unconditional previsions from qualitative judgements

- (sections 4.1, 4.4, 4.5), or from many types of quantitative judgements (section 4.6);
3. constructing probability models by refinement, marginalization or multivalued mapping (section 4.3);
 4. combining probability models from several sources (section 4.3);
 5. analysing evidence in the form of relative frequencies (sections 5.3, 5.4, 9.6);
 6. updating probability models in the light of new evidence (sections 6.4, 6.11);
 7. constructing joint probabilities from marginal and conditional probabilities (section 6.7);
 8. constructing posterior probabilities from statistical sampling models and prior probabilities, through the generalized Bayes' rule (sections 7.8, 8.4, 8.5);
 9. constructing prior or predictive probabilities from sampling models and posterior probabilities (sections 8.2, 8.3);
 10. constructing joint probabilities from several marginals and judgements of independence (section 9.3);
 11. constructing joint probabilities from judgements of permutability or exchangeability (sections 9.4, 9.5);
 12. making decisions, by constructing preferences from assessments of probabilities and utilities (sections 1.5, 3.9).

More complex assessment strategies can be built up, in practical problems, by combining the simple strategies described here.

1.7.7 Criteria for selecting assessment strategies

It is clear from the thumbtack example that several assessment strategies can be used in a single problem. There is no uniquely correct strategy. Why should You prefer one strategy or model to another? Applied statisticians have considerable expertise in building statistical models, and in recognizing that a particular type of model is appropriate to a particular combination of statistical data and theoretical evidence.¹⁴ There is much less expertise in assessing epistemic probabilities without statistical models. However, the following criteria may give some guidance in selecting an appropriate strategy and in evaluating its success.

1. The strategy should allow structural features of the evidence (such as symmetries, or evidence of independence, causal relationships or mechanisms) to be represented, through the principles of external rationality or through statistical models.
2. It should lead to greater precision in assessments of the target probabilities. It may be sensible to use several different strategies to help construct

1.7 ASSESSMENT STRATEGIES

- the same probabilities. Provided these produce consistent models, they can be combined to further increase precision (see section 4.3.8).
3. The complexity of the model should be appropriate to the importance of the problem. Constructing more detailed models may enable more of the evidence to be taken into account. Often it is necessary to idealize, simplify or ignore some of the evidence because there is no convenient strategy to make use of it. The imprecision of models should be increased to allow for the distortion.
 4. The basic probability assessments should be of types that can be made easily and accurately by human assessors. There is a large psychological literature on the ways in which humans intuitively assess probabilities, and especially on the kinds of errors they make, such as giving incorrect weight to prior probabilities and sample sizes.¹⁵ As well as indicating where assessment strategies are most needed to replace human intuition, this literature suggests 'heuristic' strategies that are used successfully and might be developed into formal methods of reasoning.
 5. Psychological experiments have shown that humans are poor at intuitively combining evidence of different types, especially prior with observational evidence.¹⁶ Where possible, strategies should decompose the total evidence and use rules of natural extension to combine assessments. It is often useful to separate 'prior' and 'observational' evidence, assess joint probabilities for the targets and the observations based only on the prior evidence, and construct target probabilities based on the combined evidence by conditioning on the observations. (See section 6.11 for discussion.) In the thumbtack example, strategies 2, 3 and 4 all use this kind of decomposition.
 6. Strategies can be evaluated on pragmatic grounds, by evaluating the outcomes of the behaviour they lead to. For example, different methods for making probabilistic weather forecasts (e.g. 'the probability of rain tomorrow in Ithaca is 0.9') might be regarded as competing assessment strategies. When applied to the same meteorological data they would produce different forecasts. They could be compared by arranging hypothetical bets between them on the basis of their forecasts, as explained in Appendices H and I.
 7. The probabilities produced by a strategy may be compared directly with actual occurrences. One criterion is that strategies should be well calibrated, in the sense that their probabilities match observed relative frequencies.¹⁷ For example, in a long series of days for which a weather forecaster announces that the probability of 'rain' is at least 0.7, the relative frequency of rain should be at least 0.7.
 8. Appropriate strategies may depend on the field of application. Cohen (1977) suggests that probability judgements made in courts of law have

a very different structure from those made in medical, commercial or statistical problems. They may reflect particular values, conventions and traditions.¹⁸

9. Appropriate strategies certainly depend on the circumstances and context of an inquiry, such as the questions of interest and contemplated actions. These factors affect not only the choice of target events, but also the precision that is required in the target probabilities. Similar evidence can lead, in different circumstances, to assessments of the target probabilities which have very different precision. In a decision problem, for example, the degree of precision that is required to determine an optimal action will depend heavily on the number of feasible actions and the precision of utility assessments.¹⁹

1.7.8 Expert systems

Assessment strategies that prove useful in a particular field of application can be embedded in computer programs, to facilitate their application in future problems. Such programs are often known as **expert systems**. Expert systems are designed to act as consultants, or to replace human specialists, in such tasks as medical diagnosis, pattern classification, systems control and engineering design.

Procedures for dealing with uncertainty and indeterminacy are an essential part of expert systems. These procedures may be based on one or more assessment strategies. The basic probability assessments and data that are needed to apply the strategies are either part of the knowledge base of the expert system, or elicited from users under the guidance of the system.²⁰ The expert system works through the assessment strategy, eliciting data or assessments from the user when needed, calculates the natural extension of all the assessments, and uses this to make inferences or decisions. The process can be structured to allow users of the system some control over the questions asked and strategies used. The system can explain its reasoning when required, by describing its strategies and data.

For example, an expert system designed to diagnose the condition of patients with specific symptoms might elicit details of symptoms from a physician, and combine these with the probabilities encoded in its knowledge base, to calculate posterior probabilities for various possible conditions of the patient. If the symptoms were uninformative then the posterior probabilities would be imprecise, and the system might be unable to make a useful diagnosis until further information was obtained about the patient.

1.7.9 Bounded rationality

There is a large literature on **bounded rationality** which emphasizes the inherent limitations of human reasoning.²¹ These limitations include

constraints on the size of short-term and long-term memory, speed of computation, and abilities to perform complicated calculations. For example, You may have difficulties in satisfying the coherence conditions or in computing natural extensions, when these involve large numbers of probability assessments.²² These constraints are especially important when there is little time available for reasoning. There may be time only to make highly simplified, approximate or imprecise assessments. Bounded rationality is a further source of imprecision in conclusions.

Although the constraints of bounded rationality are often relevant in practical reasoning, a more fundamental difficulty is that we often do not know *how* to analyse a problem, because we do not have suitable models and assessment strategies. Bodies of evidence are often very complex. For example, there is an enormous amount of evidence relevant to predicting next year's inflation rate. It would be difficult to comprehend and analyse this evidence, even if sophisticated assessment strategies and statistical models were used. If necessary we can use computers to extend our memory and computational abilities, but that is to no avail if we do not know what kinds of computations are needed to make sense of complex information.

1.7.10 Normative role of coherence

The primary purpose of assessment strategies is to *construct* probability models, but they also serve to *justify* these models, especially when they are based on principles of external rationality and objective statistical models. That is, You can justify Your probability assessments by elaborating a more detailed probability model that is related more closely to the available evidence. For example, models 3 and 4 in the thumbtack example (section 1.7.5) serve to justify the probabilities they produce.

This suggests a **normative** role for the coherence principles, in evaluating whether the inferences are reasonable by examining whether the implied inferences in the form of probabilities. The first criterion is that these probabilities are coherent. A more stringent criterion is that You can extend them in a coherent way to other events, and to various kinds of conditional probabilities, and that all these probabilities are consistent with the available evidence.²³

For example, the conclusions of statistical inference should be coherent with some judgements of sampling models and prior probabilities. (This does not require that the inferences were made by assessing prior probabilities.) If the conclusions are expressed as posterior probabilities and these are coherent with the sampling model, then they are also coherent with prior probabilities that can be constructed by natural extension. One can evaluate whether the inferences are reasonable by examining whether the implied prior

probabilities are reasonable in the light of prior evidence. Several examples of this are discussed in section 8.2.²⁴

Thus the theory of coherence can be used to evaluate the process and conclusions of reasoning.²⁵ This kind of evaluation can be extended to the general procedures that are proposed as methods of reasoning. If these methods tend to produce probabilities that are incoherent or inconsistent with evidence, that is a sign that the methods are unreasonable. For example, in Chapter 7 we will criticize some popular methods of statistical inference, including Bayesian inference from improper priors and Neyman–Pearson confidence intervals, by showing that their inferences incur sure loss in common statistical problems.

1.7.11 Conclusion

The contrast between the two types of rationality is now clear. Internal rationality, formalized through principles of avoiding sure loss and coherence, is a normative property of probability models alone, irrespective of how they were constructed. External rationality concerns the more constructive and difficult steps in the reasoning process, in which probability models are constructed through analysis of evidence. External rationality does involve the application of formal rules of coherence and natural extension, as well as other principles, but it also involves the choice of assessment strategies. This choice is unformalized, and depends on objectives and intuitive judgement as well as the available evidence. Several strategies may be reasonable in the same problem. It seems unlikely that probabilistic reasoning can ever be completely formalized, because it is implausible that the intelligence and imagination needed to devise useful assessment strategies, and the judgements needed to apply them, could be completely reduced to formal principles.

1.8 Survey of related work

There is an extensive literature on mathematical models for uncertainty and indeterminacy. A survey of this literature will be given here, to help place our work in perspective, with special attention to theories of probabilistic reasoning and concepts of imprecise probability. Some of these theories are examined more closely in Chapter 5.

1.8.1 Bayesian inference and decision

The Bayesian theory of statistical inference uses precise (additive) probability measures to model personal or logical beliefs, and emphasizes the use

1.8 SURVEY OF RELATED WORK

of Bayes' rule as a method for combining prior beliefs with statistical data. The **objective Bayesian** theory dates from a paper of Bayes (1763), and was extensively developed by Laplace (1812, 1814). Modern versions, based on a logical interpretation of probability, have been proposed by Jeffreys (1931, 1983), Carnap (1952, 1962, 1971), Jaynes (1968, 1983), Box and Tiao (1973) and Rosenkrantz (1977). Following Bayes and Laplace, these theories emphasize the use of ‘noninformative’ probability distributions to model prior ignorance. These distributions are often ‘improper’ and may produce incoherent inferences. Objective Bayesian approaches are criticized in detail in sections 5.5, 5.12 and 7.4.

A different, and currently more popular, version of the Bayesian approach is based on a personalist interpretation of probability. De Morgan (1847) suggested that probability could be used to measure personal degrees of belief, but the **subjective Bayesian** theory was developed only after Borel (1924), Ramsey (1926) and de Finetti (1931, 1937) suggested a behavioural interpretation of probability as a personal betting rate.¹ The subjective Bayesian approach has been advocated especially by de Finetti (1972, 1974, 1975), whose theory is particularly relevant to the present work because of its emphasis on coherence, Good (1950, 1965, 1983), Savage (1972a, 1981) and Lindley (1965, 1971b, 1983). Some of these approaches are outlined in section 5.7.

The basic idea of **Bayesian decision theory** is that rational decision involves maximizing expected utility, where expected utilities are calculated from precise assessments of personal probabilities and utilities. This approach is due to Ramsey (1926) and Savage (1972a). Good introductions to the theory (from different points of view) are Lindley (1971a), Jeffrey (1983) and von Winterfeldt and Edwards (1986). There are various texts that cover the Bayesian theories of inference and decision at an intermediate level; see especially DeGroot (1970) and Berger (1985a).²

1.8.2 Criticisms of Bayesian precision

It was recognized in the nineteenth century that the ‘noninformative’ prior distributions suggested by Bayes and Laplace were arbitrary and unjustifiably precise. This criticism of Bayesian methods was expressed by Mill (1843), Ellis (1843), Boole (1854), Peirce (1878) and Venn (1888).

In particular, Boole (1854, Chs. 16–21) suggested alternative methods that resulted in imprecise epistemic probabilities. He first calculated the probability of interest in terms of other precise probabilities, some known and others unknown ‘arbitrary constants’. He then obtained upper and lower probabilities by taking maxima and minima with respect to the arbitrary constants. Like most later advocates of upper and lower

probability, Boole seems to regard these as bounds on unknown, precise probabilities.³

Many criticisms of the Bayesian approach in the modern literature are concerned with the over-precision of Bayesian probabilities. Savage (1972a, Ch. 4), one of the most influential Bayesians, recognized this as a major defect of the theory. Other discussions of this issue include Keynes (1921, Ch. 3), Russell (1948, pp. 358–61), Good (1950, 1952, 1976), Fisher (1956, Ch. 2), Smith (1961), Cox (1961), de Finetti and Savage (1962), Fellner (1965), Lindley (1971a, pp. 18–26), Mellor (1971, 1980), Fine (1973, 1977a, 1983, 1988), Suppes (1974), de Finetti (1975, App. 19), Kyburg (1978), Leamer (1978, 1986), Walley and Fine (1979), Gardenfors and Sahlin (1982, 1983), Levi (1980, 1982, 1985), Berger (1984) and Einhorn and Hogarth (1985). Chapter 5 of this book is largely concerned with evaluating the arguments for and against precision.

1.8.3 Bayesian sensitivity analysis

Perhaps the simplest way to introduce imprecision into the Bayesian approach is to carry out a strict Bayesian analysis with various precise probability and utility functions. This is called **Bayesian sensitivity analysis**. In statistical inference, this produces a range of posterior probabilities, and possibly indeterminate conclusions. In decision making, it produces a range of expected utilities for each action, and possibly indecision between actions. So the practical outcome of sensitivity analysis may be similar to the outcome of our approach. There are also mathematical similarities, which will be considered throughout the book.

The clearest account of the principles of Bayesian sensitivity analysis is the paper by Berger (1984). The approach dates back at least to Boole (1854), but there has been little discussion of its principles and justification until recently. (It seems to be regarded as a minor practical adjustment to the strict Bayesian approach.) Other relevant discussions include von Mises (1942), Good (1950, 1962a, 1965, 1983), Edwards, Lindman and Savage (1963), Huber (1973), Dickey (1973), Dempster (1975) and Berger (1985a). See sections 2.10 and 5.9 for further discussion of this approach.⁴

1.8.4 Keynes (1921)

The first major effort to construct a theory of imprecise probability was made by Keynes (1921).⁵ Keynes aimed to develop an inductive logic, based on a logical interpretation of probability as a ‘degree of rational belief’.

Keynes (Ch. 3) discusses many practical examples where epistemic probabilities are not precisely determined, even in principle. He concludes

that, in general, probabilities cannot be completely ordered. Precise probabilities ‘will occur comparatively seldom’ (p. 176), only in cases where there are indivisible, equally probable alternatives.⁶ It is clear that Keynes rejects the dogma of ideal precision and sensitivity analysis interpretation: the ‘ideal’ probabilities are generally imprecise.⁷

Keynes based his mathematical theory of probability on axioms concerning a non-numerical probability relation between conclusion and evidence. He gave some examples of the calculation of upper and lower probabilities (see especially Chs. 15 and 17). Unfortunately he did not propose these, or any other numerical models, as general representations of imprecise probabilities.

The main defect of Keynes’s theory is that he says little about how probability judgements can be constructed from evidence. For Keynes, probability statements express logical relations between conclusion and evidence, but these relations are to be perceived directly through ‘intuition’ or ‘logical ability’. Keynes admits that in many cases our intuition is inadequate, and the correct logical probabilities may be unknown. As pointed out by Ramsey (1926), the intuitions involved here are highly mysterious. (Ramsey suggested a behavioural definition of probability as an improvement.)

1.8.5 Comparative probability

Since Keynes, a large literature has grown up concerning the mathematics and interpretation of imprecise, epistemic probabilities. The most popular mathematical models have been those discussed by Keynes: comparative probability orderings, and upper and lower probabilities. Most of this literature differs from Keynes in adopting a personalist (rather than logical) interpretation of probability, and a sensitivity analysis interpretation of the imprecision.

Koopman (1940a, 1940b, 1941) proposed a complicated set of axioms concerning a partial comparative probability ordering, in which comparisons are made between two hypotheses conditional on different evidence. The comparisons are apparently subjective and intuitive. Assuming that any event can be partitioned into any number of equally probable events, upper and lower probabilities can be constructed from the ordering.

Most of the work on comparative probability has used the axioms of de Finetti (1931), which include a completeness axiom, as a basis for obtaining precise probabilities.⁸ This approach is followed by Savage (1972a), DeGroot (1970), Lindley (1971a) and Jeffreys (1931, 1983). See sections 4.5 and 5.7 for discussion.

The work of Fine (1973, 1977a) on comparative probability is notable

because it forcefully rejects the dogma of ideal precision and the sensitivity analysis interpretation. Walley and Fine (1979) take a similar attitude to classificatory judgements, that specified events are probable. Many of the models studied in these papers are not coherent. Suppes (1974, 1975) suggests axioms for the ‘approximate measurement’ of beliefs through a comparative probability ordering. He uses this structure to define upper and lower probabilities that are not necessarily coherent.

1.8.6 Upper and lower probability

Borel (1943, section 3.8) suggested that upper and lower probabilities could be measured behaviourally, as betting rates on or against an event. He considers the probability that a patient will recover from illness: ‘It would moreover be natural enough for Peter, modest and prudent, to refuse to set a precise value of the probability of recovery, but merely affirm that in his judgement this probability is between 0.8 and 0.9, and that, under the circumstances, if he is forced to bet on recovery, he will demand that 0.8 be adopted, but if he is forced to bet on death, he will demand that 0.9 be adopted.’

The imprecision of personal probabilities and utilities has been emphasized in many writings of Good (especially 1950, 1952, 1962a, 1962b, 1976, 1983). In particular, Good (1962a) proposed axioms for upper and lower probabilities. These are interpreted through a ‘black box’ model of probabilistic reasoning, which is a version of the sensitivity analysis interpretation. This model is discussed in section 5.9.

The fundamental principles of ‘avoiding sure loss’ and ‘coherence’ were introduced, in an informal way, in the important paper of Smith (1961). Smith interpreted upper and lower probabilities as personal betting rates, and deduced their mathematical structure by applying the coherence principles. He also extended the principles to statistical inference and decision making. He presented his work as a generalization of the subjective Bayesian theory to admit imprecision.

Williams (1976) gave axioms to characterize coherence of upper and lower probabilities and previsions, again interpreted as personal betting rates, and generalized Smith’s result that coherent lower probabilities are lower envelopes of precise probability measures. (That result suggests a sensitivity analysis interpretation, although it is not clear whether Smith or Williams favoured that interpretation.⁹) An unpublished paper of Williams (1975a) contains important ideas on conditional probability. The mathematical theory in this book is based on the ideas suggested by Smith and Williams.

Fishburn (1964, p. 192) regarded upper and lower probabilities, elicited directly or via comparative judgements, as upper and lower bounds on ‘true’ personal probabilities. Similarly, in the theory of Fellner (1961, 1965), a

person ‘discloses through his behaviour only *ranges* within which his true probabilities are located’. (It is unclear what is meant by ‘true probabilities’.) The imprecision in probabilities is ascribed to the psychological instability of beliefs and to disagreement between persons. Various methods for modelling this kind of disagreement through imprecise probabilities were defined and compared by Walley (1982). The applications described by Thorp, McClure and Fine (1982), concerning long term forecasting of electricity costs, and Walley and Campello de Souza (1986), concerning evaluation of the economic viability of solar heating systems, used upper and lower probabilities to reflect the disagreement amongst expert opinions.

Leamer (1986), using ideas from the economic theory of auctions, has studied operational schemes for eliciting upper and lower probabilities as selling and buying prices for lottery tickets. He assumes that underlying beliefs are determinate, but they may be unstable or difficult to elicit.

Dempster (1966, 1967a, 1968) suggested a new approach to statistical inference, based on a special type of coherent lower probability called a belief function, which can be generated from a precise probability measure through a multivalued mapping. The theory and applications of belief functions have been developed by Shafer (especially 1976, 1981a, 1982a). This theory is especially appealing because it emphasizes the construction of belief functions from simple judgements, and because it does not rely on the Dogma of Ideal Precision. Unconditional belief functions are coherent, and therefore compatible with our theory. However, Dempster’s rule for defining conditional belief functions can produce a sure loss. The Dempster–Shafer theory is discussed in section 5.13.¹⁰

1.8.7 Preference and decision

Another approach is based on axioms for partial preference orderings of gambles.¹¹ (Compare with our axioms in section 3.7.) Buehler (1976) characterized the partial preference orderings that avoid sure loss. Giles (1976) presented an operational theory of subjective probability, based on an agent’s commitments to accept specific gambles from an opponent. Giles imposed coherence axioms concerning sets of acceptable commitments (essentially, these define a partial ordering of gambles), and showed that these can be represented in terms of a ‘risk function’ (essentially, a coherent upper prevision). Giron and Rios (1980) also proposed axioms for partial preference orderings, and used these to suggest ‘quasi-Bayesian’ methods for statistical inference and decision.

Wolfson and Fine (1982) and Wolfson (1979) suggested methods of inference and decision based on upper and lower expectations. They require specification of precise probabilities but admit imprecision through ‘lack of

confidence' in these. Kmietowicz and Pearman (1981) describe several methods for making decisions using imprecise probabilities.

There has been some work in experimental psychology to investigate how people do make decisions when there is indeterminacy (or 'ambiguity') about relevant states of affairs. See, in particular, Ellsberg (1961), Fellner (1961, 1965), Becker and Brownson (1964), Yates and Zukowski (1976), Curley and Yates (1985), Hogarth and Kunreuther (1985), Einhorn and Hogarth (1985, 1986), and Budescu *et al.* (1988). These papers report experimental results on the ways in which indeterminacy affects choices. They also suggest descriptive models to account for these results. As indeterminacy is a characteristic of most practical reasoning and decision, it is surprising that there have not been more experimental studies, especially to investigate what kinds of inference strategies people use to cope with indeterminacy.

1.8.8 Epistemological theories

There have been several attempts to establish a theory of logical probability which admits imprecision. Kyburg (1961, 1970, 1974a, 1983a) has presented a detailed epistemological theory of upper and lower probability which focuses on the structure of bodies of evidence and the beliefs they warrant. Kyburg argues that the evidence often includes knowledge that relative frequencies belong to specified intervals. Central to his theory are rules for constructing upper and lower probabilities for the hypothesis that an individual has some property, by combining different intervals of relative frequencies obtained from different reference classes to which the individual belongs. Kyburg's rules for conditioning appear to violate our principles of coherence.

The work of Levi (1980) is valuable as an integrated theory of knowledge, evidence, beliefs, values and decisions that pays due attention to indeterminacy. (See also Levi, 1967, 1974, 1982, 1985.) Like Bayesian sensitivity analysts, Levi uses classes of precise probability measures to represent indeterminate beliefs, although he strongly rejects the sensitivity analysis interpretation and the dogma of ideal precision. Rational beliefs are modelled by the convex class of all probability measures that cannot be 'ruled out as impermissible' relative to given evidence, by applying basic rationality axioms plus certain contextual considerations. Levi gives more emphasis than we do to the precise probability measures in this class. (For instance, he formulates his rationality axioms and rules for conditioning in terms of them.) Apart from this difference in emphasis, Levi's theory of probability appears to be compatible with ours. There are, however, differences between the two approaches to decision making, especially when both probabilities and utilities are imprecise. Like sensitivity analysts, Levi

1.8 SURVEY OF RELATED WORK

permits only those actions that are optimal with respect to some pair of precise probability and utility functions.¹² When there are several such actions, he advocates using a minimax rule to resolve the indecision.

Similarly, Gardenfors and Sahlin (1982, 1983) model indeterminate beliefs by a class of precise probability measures (those measures whose 'reliability' exceeds a certain level), and use another minimax principle for making decisions.

1.8.9 Frequentist interpretations

A frequentist interpretation of upper and lower probability has been used in studies of the robustness of statistical procedures, notably by Huber and Strassen (1973) and Huber (1973, 1981). Huber (1981, Ch. 10) presents some fundamental results concerning coherence of upper and lower probabilities and expectations.

Walley and Fine (1982) suggest that the imprecision of upper and lower probabilities can model physical indeterminacy in a phenomenon, and not merely lack of knowledge about underlying chances. On a limiting frequency interpretation, upper and lower probabilities can be identified with the upper and lower limits of relative frequencies that result from independent repetitions of an experiment. Upper and lower probability models defined in this way are always coherent. These aleatory interpretations will be discussed in section 7.2, when statistical sampling models are introduced.

A quite different type of upper and lower probability has been advocated by Kumar and Fine (1985), Grize and Fine (1987), Papamarcou and Fine (1986) and Papamarcou (1987), to model stationary processes with unstable time averages. Fine (1988) is a good introduction to this work.

1.8.10 Uncertainty in expert systems

Expert systems are computer programs that encode the judgements and reasoning procedures of human experts. They are designed to aid or replace humans in specific tasks. Introductions to expert systems include Duda and Shortliffe (1983), Forsyth (1984), Waterman (1986) and Stock (1987).

Pearl (1988), and Lauritzen and Spiegelhalter (1988), have defended subjective Bayesian models for uncertainty in expert systems, and developed computational techniques for propagating probabilities through causal networks. However, it is widely recognized that expert systems need to deal with incomplete, inconsistent and qualitative information, and that there are often difficulties in assessing the many precise probabilities needed to apply Bayesian methods. Various other ways of modelling uncertainty have been applied in expert systems. Spiegelhalter (1986) is a good survey of these

methods.¹³ Many of the models for uncertainty, as well as the rules and procedures used in these systems, are *ad hoc* and seem to be chosen mainly because of their computational simplicity. Consequently, the meaning of assessments and conclusions is unclear. The expert systems which use **fuzzy logic** (discussed in section 5.11) are one example; see Zadeh (1983) and Prade (1985).

Another well-known example is the method of **certainty factors** used in the expert system MYCIN, which can be used by physicians to diagnose (and prescribe treatments for) bacterial infections of the blood. This system is described by Shortliffe and Buchanan (1975), Shortliffe (1976), Buchanan and Shortliffe (1984), Spiegelhalter and Knill-Jones (1984). MYCIN uses separate measures of belief and disbelief in a hypothesis. The behavioural meaning of these measures is unclear (they appear to conflate probability with utility), and the rules for combining measures and for combining evidence appear to be unjustified, except under strong assumptions of independence; see Adams (1976).¹⁴

The expert system INFERNO, described by Quinlan (1983), is based on upper and lower probability. The inference rules of the system always produce valid conclusions, but stronger conclusions could be obtained, in general, by applying rules of natural extension.¹⁵

1.8.11 Other models for uncertainty

Cohen (1977) has argued that the standard theory of additive probability cannot adequately model important features of legal reasoning. He introduces **Baconian probabilities** to measure ‘degree of provability’, and uses these to model (and thereby defend) six simple types of legal argument. A basic mathematical property of Baconian probabilities is that a hypothesis and its negation cannot both have positive probability. (The logic of surprise proposed by Shackle, 1952, 1961, has similar properties.) Baconian probabilities can be regarded as a special type of coherent lower probability, but (whatever their merits as models for legal reasoning) they are insufficiently general to model most types of uncertainty.¹⁶

The theory of **fuzzy sets** was introduced by Zadeh (1965, 1978) as a way of dealing with the vagueness and ambiguity of ordinary language. Some authors have argued that this kind of ambiguity is quite different from probabilistic uncertainty, and that fuzzy and probabilistic methods are applicable to different problems. Others, notably Watson, Weiss and Donnell (1979) and Freeling (1980), have defined ‘fuzzy probabilities’ and constructed a ‘fuzzy decision theory’. This approach is evaluated in section 5.11.

1.8.12 Types of imprecise probabilities

It is clear that a variety of mathematical models have been proposed to represent uncertainty and indeterminacy. However, the differences between

these models are not as great as they may appear. Almost all of the mathematical models surveyed in sections 1.8.1 to 1.8.9 are special cases of, or are broadly compatible with, the models developed in this book. In particular, we will study the following types of mathematical model, for which we use the generic term **imprecise probabilities**¹⁷:

1. upper and lower previsions (defined in section 2.3);
2. upper and lower probabilities (section 2.7);
3. classes of additive probability measures (or linear previsions) (sections 2.8, 3.3);
4. classes of desirable gambles (section 3.7, Appendix F);
5. partial preference orderings of gambles (section 3.7);
6. classifications of probable events (section 4.4);
7. comparative probability orderings of events (section 4.5);
8. belief functions (section 5.13).

Most of our mathematical theory will be presented in terms of **lower previsions**, which determine upper previsions. Of the other models, 2 and 8 are special types of lower previsions, and 6 and 7 determine lower previsions in a natural way. The remaining models 3, 4, 5 are mathematically equivalent to lower previsions, in the sense that there are one-to-one correspondences between models of these types when suitable coherence axioms are imposed (see section 3.8). So if You can assess any of the listed models, You can use them to generate lower previsions. Because of that, the choice of lower previsions as our basic model should not be controversial.¹⁸

CHAPTER 2

Coherent previsions

In this chapter we present the basic theory of coherent upper and lower previsions. The main aims are to:

1. define ‘avoiding sure loss’ and ‘coherence’, and justify them as rationality criteria;
2. derive the basic properties of coherent previsions;
3. examine the special cases of upper and lower probabilities, linear previsions and additive probabilities;
4. present some useful examples of coherent upper and lower previsions;
5. discuss fundamental issues of interpretation and objections to the behavioural interpretation we favour.

The theory is developed in terms of a mathematical model $(\Omega, \mathcal{K}, \underline{P})$. Here Ω is a set of possible states of affairs whose interpretation is discussed in section 2.1. \mathcal{K} is a class of gambles on which a real-valued function \underline{P} , called a lower prevision, is defined. A gamble X is a bounded real-valued function on Ω which is interpreted as an uncertain reward. The rewards $X(\omega)$ are measured in units of linear utility, such as the ‘probability currency’ constructed in section 2.2.

In section 2.3 we introduce lower previsions \underline{P} as models for beliefs about Ω . Our behavioural interpretation is that You are disposed to pay any price less than $\underline{P}(X)$ for the gamble X . When \mathcal{K} is a linear space of gambles, coherence of \underline{P} is characterized by three axioms, which hold for all X and Y in \mathcal{K} and positive constants λ :

- (P1) $\underline{P}(X) \geq \inf X$,
- (P2) $\underline{P}(\lambda X) = \lambda \underline{P}(X)$
- (P3) $\underline{P}(X + Y) \geq \underline{P}(X) + \underline{P}(Y)$.

The behavioural interpretation of \underline{P} is used to justify these axioms. The conjugate upper prevision \bar{P} is defined by $\bar{P}(X) = -\underline{P}(-X)$ and can be interpreted as an infimum selling price for X . Simple examples of coherent lower previsions are the linear previsions, which are characterized by the

COHERENT PREVISIONS

self-conjugacy property $\underline{P}(X) = \bar{P}(X)$, and the vacuous previsions, defined by $\underline{P}(X) = \inf X$ and $\bar{P}(X) = \sup X$.

The criteria of avoiding sure loss and coherence are defined for an arbitrary domain \mathcal{K} in sections 2.4 and 2.5. A lower prevision \underline{P} avoids sure loss when there is no finite combination of transactions in which You buy gambles for prices less than $\underline{P}(X)$ which is certain to produce a net loss. Coherence is violated just when there are some such transactions whose net result is that You pay more than $\underline{P}(X_0)$ for some gamble X_0 in \mathcal{K} . Coherence implies avoiding sure loss, and reduces to the axioms P1–P3 when \mathcal{K} is a linear space. The basic properties of coherent lower previsions are listed in section 2.6. From a given set of coherent lower previsions one can construct new coherent lower previsions by forming lower envelopes, convex combinations, or limits of sequences.

Upper and lower probabilities are defined in section 2.7 as special cases of upper and lower previsions, for which the domain \mathcal{K} contains only zero-one valued gambles. The main reason for regarding previsions as more fundamental than probabilities is that coherent lower probabilities may have many different extensions to coherent lower previsions. (That is illustrated by example 2.7.3.) In section 2.8, linear previsions and additive probabilities are obtained as special types of lower previsions and lower probabilities. When \mathcal{K} is a linear space containing constant gambles, the linear previsions are just the positive linear functionals with unit norm. Additive probabilities defined on a field of events are characterized by the standard axioms for finitely additive probability.

Various examples of coherent upper and lower previsions are given in section 2.9, including models for pari-mutuel betting systems, insurance and capital gains tax, inner and outer Lebesgue measure, ‘uniform distributions’ on the real line and positive integers, and zero-one valued lower probabilities generated by filters of sets.

Some possible extensions of the minimal behavioural interpretation of lower previsions are discussed in section 2.10. We especially consider operational interpretations (\underline{P} represents actual gambling commitments), exhaustive interpretations (\underline{P} exhaustively models Your dispositions to accept gambles), and sensitivity analysis interpretations (\underline{P} is the lower envelope of a class of linear previsions which represent hypothetical ‘ideal’ belief-states). None of these interpretations seems to be generally applicable. Finally, in section 2.11 we discuss some objections to the behavioural interpretation. The objections concern the use of prior probabilities in statistical inference, various notions of objectivity, distinctions between types of uncertainty, the emphasis on behaviour and avoiding sure loss, the need for other rationality criteria, probability assessment, and the meaning of beliefs about unobservable states.

2.1 Possibility

Probabilities and previsions will be constructed on a set Ω called the **possibility space**. Briefly, Ω is a set of **possible states** ω which describe some aspect of the world on which interest is focused. Often Ω will be the set of possible outcomes of a well-defined experiment or observation. (In classical probability theory, Ω is called the sample space.) In statistical problems, we will take the possible states to be all possible ordered pairs of parameter values and statistical data.

The states ω of the possibility space Ω are required to be mutually exclusive (it is not possible that more than one state occurs) and exhaustive (it is not possible that none of the states occurs). Disjunctions of possible states will be called **events** and identified with subsets of Ω in the usual way.

Two issues will be discussed briefly in the following sub-sections. First, what sort of things are the states ω ? Second, what does it mean to say that a state is possible?¹ The interpretation of possibility affects the meaning of ‘mutually exclusive’ and ‘exhaustive’, and it will also affect the interpretation of avoiding sure loss and coherence. (An event is **sure** or **certain** just when it is not possible that the event does not occur.)

2.1.1 States of affairs²

The states ω may represent the outcomes of experiments or observations, events or occurrences, or any conceivable facts about the world. In particular, ω need not represent an ‘event’ in the usual sense, e.g. ω might refer to the ‘state’ that a specified person is over 6 feet tall. For our purposes, the only essential requirement is that a unique element of Ω , which we will call the ‘true state’, is determined in some way. (Of course, in the problems of interest, You are uncertain about which state is the true state.) Usually the states will be regarded as mutually exclusive hypotheses concerning some external reality, and the true state is the one that correctly describes that reality. But the ‘true state’ might be distinguished in some other way; it might be selected by a person or group, for example.

Our theory is formulated in terms of gambles, which are interpreted as uncertain rewards whose value depends on the true state. For such a formulation to be operational, it is necessary that the true state be determinable by some clearly defined verification procedure, by some specified future time. De Finetti (1974, 1975) restricts his theory to deal only with states of affairs that are observable in this way.³ That is a drastic restriction on the scope of an epistemic theory. For example, it rules out theoretical states like ‘this coin is biased towards heads’ because no finite verification procedure can be given. But it seems meaningful to regard ‘bias towards heads’ (a tendency to fall heads more often than tails) as a physical state

2.1 POSSIBILITY

of the coin that is or is not present. It also seems meaningful to have beliefs about whether the coin is biased towards heads, irrespective of our ability to determine the answer.

The issue of observability is discussed in section 2.11.9. Briefly, our approach is to admit theoretical (unobservable) states as well as observable ones. Your beliefs about unobservable states can be measured indirectly, through their effects on beliefs about observable states such as the outcomes of future coin-tosses, or more directly, through Your attitudes to hypothetical gambles which could be settled only if the unobservable state could be determined. Thus we admit all possibility spaces for which a unique state is distinguished as ‘true’, even if You may never know which is the true state. The coherence conditions are, however, somewhat easier to interpret, and more compelling, when they concern spaces Ω whose states are observable by some specified time.

2.1.2 Concepts of possibility

We will outline four interpretations of ‘possibility’. First, a state of affairs ω is **epistemically possible** (for You) when it is logically consistent with Your available information. Epistemic possibility is relative to Your information, but is ‘objective’ in the sense that it is determined merely by this body of information and does not depend on subjective judgement. This concept of epistemic possibility is discussed in some detail by de Finetti (1974, Ch. 2), but seems problematical for a theory (like de Finetti’s) which does not require that a ‘body of information’ be precisely defined; for then the logical implications of this information may be ambiguous. Even if we require that a body of information be completely specified in some formal language, so that its logical implications are well defined, You may not know whether particular states (concerning unsettled mathematical conjectures, for example) are epistemically possible, because of Your limitations in logical and mathematical analysis.⁴

Both these difficulties can be avoided by adopting a weaker notion of apparent possibility. A state is **apparently possible** (for You) unless You have determined that it is logically inconsistent with Your available information. A set of apparently possible states may change without any change in Your information, through further specification or analysis of this information. All the integers from 0 to 9 are apparently possible (for me, at this time) as the tenth digit in the decimal expansion of π , but only one integer is epistemically possible (consistent with my knowledge of basic mathematics). I could determine the integer by doing some calculation, thereby reducing the set of ten apparent possibilities to one.

Some apparently or epistemically possible states may be negligible. As a simple example, consider a toss of a coin performed in the usual way.

The usual possibility space is $\Omega_1 = \{H, T\}$, where H denotes the outcome that the coin lands with ‘head’ up, and T that it lands with ‘tail’ up. But other outcomes, such as the coin landing on its edge, are logically consistent with our background knowledge about coin-tossing. To obtain an exhaustive space we must include such bizarre occurrences as the coin breaking into pieces on landing, or disappearing down a crack in the floor opened up by an earthquake. These are epistemic possibilities, although they are negligible in practice.⁵

Possibility spaces like Ω_1 can be admitted by strengthening the definition of possibility. One approach is to ignore epistemically possible states in discrete spaces Ω which have zero probability. More generally, say that an epistemically possible state ω is **practically possible** if (in a suitable topology) every open set containing ω has positive upper probability. This concept of practical possibility is not much stronger than epistemic possibility. In fact, under weak regularity conditions, the set Ω of all practically possible states will have lower probability one, meaning that You are prepared to bet at any odds that Ω will occur. You might be willing to bet on the coin landing on its edge at sufficiently generous odds, and therefore judge this outcome to be practically possible, but still wish to restrict attention to the space Ω_1 .

To do so, You could allow Ω to be any set of **pragmatic possibilities** that are regarded as sufficiently important to be included in a model. The set Ω may not be practically or epistemically exhaustive, and may have probability less than one. A probability model is then regarded as conditional on Ω , and elicited through Your attitudes to gambles that are called off if Ω fails to occur.⁶ For example, You bet on ‘heads’ with the understanding that the bet is called off if the coin lands on its edge. It might be argued that most probability models are ‘conditional’ in this way, in that they are based on simplifying assumptions which eliminate states that are epistemically possible but unimportant or unlikely.

More seriously, Ω may be far from exhaustive when we are unable, because of our limited understanding of some phenomenon, to imagine what might happen. We are unable, for example, to identify all the possible ways in which a nuclear reactor may break down. In some cases we may have little confidence that our possibility space Ω includes all the serious possibilities.

2.1.3 Refinement of Ω

It is important that the states ω be sufficiently detailed descriptions of the domain of interest, so that Your beliefs can be adequately expressed through Your attitudes to gambles defined on Ω . Here ‘adequacy’ depends on the particular task (e.g. decision problem) for which Your beliefs need to be

modelled.⁷ Since our theory allows Your beliefs about any domain to be indeterminate, there is nothing to be lost, except perhaps in comprehensibility or mathematical tractability, by taking the states to be more detailed than needed. In any case, the space Ω can be reformulated at any stage of an analysis, for example by refining elements of Ω into more detailed states or by extending Ω to include new pragmatic possibilities. Indeed, such reformulation will often be necessary in complex problems. One way of looking at this is that Ω is a mathematical construct that evolves in time; Ω_t is the smallest space in terms of which the probabilities assessed up to time t can be fully expressed. Various sorts of reformulation of Ω are discussed in section 4.3.

We stress that Your beliefs, even those expressed with respect to a fixed Ω , are regarded as evolving in time. All probability models refer to a particular time, though explicit reference to time will be suppressed except when we discuss updating of beliefs.

2.1.4 Random quantities

De Finetti (1974, Ch. 2) prefers to formulate his theory in terms of **random quantities** without referring to an underlying possibility space Ω . Random quantities are simply uncertain quantities whose possible values are real numbers. There is an obvious correspondence between the two formulations. Given a possibility space Ω , random quantities can be identified with real-valued functions (gambles) defined on Ω . Conversely, given a collection of random quantities, the possible states can be identified with the possible combinations of values of all the random quantities. (In practice, Ω might be constructed in just this way from the random quantities referred to in elicitation.) Because of this correspondence, the two formulations are mathematically equivalent.

De Finetti prefers to work with random quantities partly because of the temptation, reinforced by classical probability theory, to regard Ω as fixed or unique in a particular problem. In constructing and developing probability models one often needs to introduce new random quantities. This corresponds to refinement of the possibility space Ω , but it seems psychologically easier to accept the new random quantities than to accept the corresponding modification of Ω . One might also argue in favour of random quantities that they have a primary role in elicitation of beliefs, and that much of the literature on stochastic processes and statistics is directly concerned with random variables and their joint distribution rather than with an underlying sample space.⁸ Nevertheless, we prefer a formulation in terms of possibility spaces because of its heuristic advantages (familiarity from the classical theory of probability), the useful identification of events with sets, its concreteness in specifying the possible states that are under

consideration, and because the formulation of an appropriate space Ω is often a first step in practical problems.

2.2 Probability currency

In the next section, lower and upper previsions will be introduced as buying and selling prices for gambles defined on a possibility space Ω . A gamble is a bounded real-valued function defined on domain Ω . A gamble should be interpreted as a reward whose value depends on the uncertain state ω . If You accept the gamble X , then at some later time the true state ω will be determined and You will receive the reward $X(\omega)$, in units of utility. In this section we construct a simple linear utility scale for rewards. The properties of this scale will then serve to motivate and justify the coherence conditions introduced in the following sections. First we list the basic notation used throughout the book.

2.2.1 Notation

Ω denotes a possibility space, ω denotes an arbitrary element of Ω . Gambles (X, Y, Z) are bounded functions from Ω to \mathbb{R} (the set of real numbers). Write $\inf X$ and $\sup X$ for the infimum (greatest lower bound) and supremum (least upper bound) of $\{X(\omega) : \omega \in \Omega\}$. (Here \in means ‘is an element of’.) Both $\inf X$ and $\sup X$ are finite for all gambles X . Let $\mathcal{L}(\Omega)$ denote the set of all gambles on Ω ; this will be denoted by \mathcal{L} when there is no ambiguity about the underlying space Ω . We will use i, j, k, m, n to denote integers. Greek letters $\alpha, \beta, \delta, \epsilon, \zeta, \eta, \lambda, \mu, \rho, \sigma, \tau$ will be used to denote real numbers, and also to denote constant gambles, e.g. the gamble λ takes the constant value $\lambda(\omega) = \lambda$ for all $\omega \in \Omega$. For all $\lambda \in \mathbb{R}$ and $X \in \mathcal{L}$, λX is the gamble defined by $(\lambda X)(\omega) = \lambda X(\omega)$. For all $X \in \mathcal{L}$ and $Y \in \mathcal{L}$, $X + Y$ is the gamble $(X + Y)(\omega) = X(\omega) + Y(\omega)$. So $X + \lambda$ is the gamble $(X + \lambda)(\omega) = X(\omega) + \lambda$. The set of all gambles \mathcal{L} is a real linear space under these operations (see Appendix D). Finally, write $X \geq Y$ to mean $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$.

Note that gambles are required to be bounded. We do not consider the unbounded loss functions, such as quadratic loss, that are commonly used in statistical decision theory. (These are equivalent to ‘gambles’ that are bounded above but unbounded below.) Unbounded utility functions do not seem to be realistic, except perhaps in problems of ‘life and death’ (or more serious issues). Certainly the utility functions relevant to ordinary problems of statistical inference and decision are bounded.¹ While our theory could be generalized to admit unbounded gambles, to do so would introduce technical complications and would mislead the reader by suggesting that sensible utility functions might, in practice, be unbounded.

2.2 PROBABILITY CURRENCY

For many people the utility of a monetary reward is roughly a linear function of its monetary value, as long as the amounts of money involved remain ‘within appropriate limits’.² In that case the rewards $X(\omega)$ of a gamble X could be interpreted as amounts of money. This may be the most convenient interpretation for readers who have some experience of evaluating monetary gambles but find the concept of utility somewhat abstract. On the other hand, people do not always judge the utility of money to be linear in monetary value, and it is not a requirement of rationality that they should do so. We will therefore show how to construct an effective utility scale in which ‘utility’ has a concrete interpretation. In fact, we take the rewards to be tickets in a lottery, and we can identify the utility of a reward with the number of tickets gained.³

2.2.2 Probability currency⁴

Consider a lottery with a single valuable prize and a large number of tickets. It is known to You that each ticket has the same chance of winning the prize, and that the outcome of the lottery is independent of the states ω that concern You and also independent of any action (accepting or rejecting gambles) You might take. Thus we assume the existence of an extraneous randomizing device which selects the winning ticket. Your chance of winning the prize is then equal to the proportion ρ of lottery tickets that You hold. Thus ρ is a known, precise, aleatory probability. For simplicity, we will assume that ρ can take all values between 0 and 1. (Tickets are infinitely divisible.)⁵

Suppose You are interested in some possibility space Ω . A gamble X defined on Ω is interpreted as an uncertain reward in **probability currency**. That is, if You accept the gamble X then, after ω is observed, You will receive a number of lottery tickets proportional to $X(\omega)$, which changes Your chance of winning the prize from ρ_0 to $\rho_1(\omega) = \rho_0 + \alpha X(\omega)$. (Of course, if $X(\omega)$ is negative then You lose lottery tickets and Your chance of winning the prize is reduced.) Here the positive proportionality constant α , which determines the ‘stake’ of the gamble X , must be chosen independently of ω but sufficiently small to ensure that $0 < \rho_1(\omega) < 1$. That is possible since X is bounded; it suffices to choose $\alpha < \min\{\rho_0, 1 - \rho_0\}/\sup|X|$. If many gambles X_1, X_2, \dots, X_n are considered simultaneously, α must be chosen small enough to ensure that $0 < \rho_1(\omega) < 1$ irrespective of which gambles are accepted and which ω occurs, where now $\rho_1(\omega) = \rho_0 + \alpha \sum X_j(\omega)$ and the sum is over all gambles X_j that are accepted. Subject to these constraints, the argument below establishes that the desirability of a gamble should not depend on what stake α is chosen.

So a gamble X is interpreted as a random reward of αX in probability

currency, or (equivalently) as a reward of a random number $M\alpha X$ of lottery tickets, where M is the total number of tickets. With this interpretation, we can argue for the following constraints on the set of gambles that You judge desirable. In saying that a gamble X is ‘desirable’ to You, we mean that You have a disposition to accept X whenever it is offered to You. When X is not desirable, You do not necessarily have a disposition to reject it – You may be undecided about whether to accept or reject. The following conditions are proposed as rationality constraints on Your set of desirable gambles. They require, in effect, that the scale in which rewards $X(\omega)$ are measured behaves like a linear utility scale. When the rewards are lottery tickets, Your utility must be a linear function of the number of tickets You hold.⁶

2.2.3 Axioms for desirability

- (D0) If $\sup X < 0$ then X is not desirable.
- (D1) If $\inf X > 0$ then X is desirable.
- (D2) If X is desirable and λ is a positive real number then λX is desirable.
- (D3) If X and Y are each desirable then $X + Y$ is desirable.

2.2.4 Justification⁷

We now attempt to justify these axioms under the interpretation of X as an uncertain reward of αX in probability currency, where the positive stake α satisfies the above constraints.

- (D0) If $\sup X < 0$ then $\sup(\alpha X) = \alpha \sup X < 0$, and by accepting X You will certainly reduce Your chance of winning the valuable prize. Therefore X should not be desirable.
- (D1) If $\inf X > 0$ then $\inf(\alpha X) > 0$, and accepting X will certainly increase Your chance of winning the prize.
- (D2) Compare (a) a reward of αX in probability currency, with (b) a reward of αX in probability currency provided an extraneous random event C with known positive chance β (independent of ω) occurs, otherwise zero reward. Since (b) yields zero unless C occurs, it should be judged desirable if and only if it is desirable conditional on the occurrence of C . Conditional on C , (b) is equivalent to (a). So (b) should be desirable if and only if (a) is desirable. But, under each state $\omega \in \Omega$, accepting (b) increases Your chance of winning the prize by $\beta\alpha X(\omega)$, hence is equivalent to a reward of $\beta\alpha X$ in probability currency. This shows that the reward $\beta\alpha X$ should be desirable if and only if αX is desirable, for every β such that $0 < \beta \leq 1$. Thus the

desirability of X should not depend on the stake α that is used to scale rewards, and D2 follows.⁸

(D3) Suppose that X and Y are each desirable. Introduce an extraneous random experiment in which event C has chance $\frac{1}{2}$ (independent of ω); e.g. C is the event that a fair coin lands ‘heads’. Consider the compound gamble Z which yields αX in probability currency if C occurs, and yields αY otherwise, for some positive stake α . Then Z is desirable if C occurs (since X is), and desirable also if C fails to occur (since Y is), so Z should be unconditionally desirable. Under state ω , accepting Z increases Your chance of winning the prize by $\frac{1}{2}\alpha X(\omega) + \frac{1}{2}\alpha Y(\omega)$, so Z is equivalent to the reward $\frac{1}{2}\alpha(X + Y)$ in probability currency. By D2, the gamble $X + Y$ should be desirable.⁹

The axioms D0–D3 can be justified in a similar way when rewards are expressed in any linear utility scale.¹⁰ Linearity of utility is reflected in axioms D2 and D3, which imply that the set of desirable gambles forms a convex cone in $\mathcal{L}(\Omega)$.

2.3 Upper and lower previsions

We can now introduce the probabilistic models on which our theory is based: lower previsions and their corresponding upper previsions. A **lower revision** \underline{P} is a real-valued function defined on some class of gambles \mathcal{X} . Here \mathcal{X} , which is called the **domain** of \underline{P} , is an arbitrary subset of the class $\mathcal{L}(\Omega)$ of all gambles on the possibility space Ω . When it is necessary to identify the underlying possibility space Ω and domain \mathcal{X} , we will refer to the triple $(\Omega, \mathcal{X}, \underline{P})$ as a lower revision.¹

2.3.1 Behavioural interpretation

The gambles X in \mathcal{X} should be regarded as uncertain rewards, expressed in units of utility such as the probability currency constructed in section 2.2. The lower revision \underline{P} then models Your attitudes (behavioural dispositions) concerning these gambles. Specifically, our behavioural interpretation of the lower revision $(\Omega, \mathcal{X}, \underline{P})$ is that, for each X in \mathcal{X} , You are currently willing to pay any price strictly less than $\underline{P}(X)$ for the gamble X . (Here both X and $\underline{P}(X)$ are measured in the same units of utility.) So the lower revision $\underline{P}(X)$ can be regarded as a **supremum buying price** for the gamble X : $\underline{P}(X)$ is the supremum price μ for which it is asserted that the gamble $X - \mu$ is desirable to You.

This behavioural interpretation of \underline{P} is minimal in the following respects (most of these comments will be elaborated in section 2.10):

1. We do not require that the primary purpose of the model \underline{P} is to describe Your attitudes to gambles, merely that the model has the stated implications concerning gambles. It may have other implications. The model \underline{P} need not be given a personalist or subjective interpretation. It might have a logical interpretation, as a summary of the buying prices for gambles that are justified on the evidence.
2. We do not require that \underline{P} be constructed or elicited through direct assessment of buying prices for gambles, merely that the values $\underline{P}(X)$, however constructed, can then be interpreted as supremum buying prices.²
3. We do not require that \underline{P} be elicited through an operational measurement procedure, in which gambles are actually bought and sold at prices acceptable to You. (In one kind of operational procedure You would set betting odds, represented by \underline{P} , as bookmakers do.)
4. It is claimed only that You are disposed to pay any price less than $\underline{P}(X)$ for X . We do not rule out the possibility that You are disposed to pay more than the specified value $\underline{P}(X)$ for X . If so, the model \underline{P} is an incomplete or non-exhaustive description of Your beliefs. We expect that, in practice, models will often fail to be exhaustive, due to incomplete modelling and elicitation.
5. Even if You do not have a disposition to pay more than $\underline{P}(X)$ for X , You may have no contrary disposition; there may be real indeterminacy in Your beliefs. In that case, when given the option of paying price μ (greater than $\underline{P}(X)$) for X , Your choice is not determined. You may decide either way. Indeed, we do not assume that You have any non-trivial behavioural dispositions, and nor will this be required by our coherence axioms. (The lower revision $\underline{P}(X) = \inf X$ is then said to be **vacuous**.)
6. We do not require \underline{P} to be constructed as the lower envelope of some class \mathcal{M} of additive probability measures, as in Bayesian sensitivity analysis, although many of our examples will be constructed in this way. Whereas \underline{P} has a direct behavioural interpretation, the interpretation of \mathcal{M} is more problematical, and we therefore regard lower revisions as more fundamental than classes of probability measures.
7. The behavioural interpretation of \underline{P} refers only to the atypical or artificial situations in which Your utility function is precise. (Gambles are just precise linear utility functions.) This restricted interpretation suffices in developing a theory of probability and statistical inference, but it must be extended to imprecise utilities in order to construct a general theory of decision making.
8. No restrictions are placed on Ω or \mathcal{K} . The domain \mathcal{K} may be any class of gambles on Ω . When \mathcal{K} is a class of indicator functions of events, for example, the lower revision $\underline{P}(A)$, called the **lower probability** of A , is a

supremum betting rate on the event A . In this section we define coherence of \underline{P} in the special case where \mathcal{K} is a linear space of gambles, but the definition will be generalized to an arbitrary domain \mathcal{K} in section 2.5. The special case of lower probabilities is discussed in section 2.7.

Because of its minimal requirements, the behavioural interpretation of \underline{P} outlined above should be acceptable to scholars with widely differing views about probability. Non-Bayesians may deny (as we do) the Bayesian dogma that supremum buying prices for gambles should always be equal to infimum selling prices, but they can hardly deny the much weaker claim that people can assess some buying and selling prices for gambles (especially as these prices may be vacuous!). Despite its minimality, the behavioural interpretation is specific enough to guide and to justify the theory of coherent revisions developed in this book. In our view, the more specific interpretations mentioned above (logical, operational, exhaustive, Bayesian, and Bayesian sensitivity analysis) should be regarded as extensions of the minimal behavioural interpretation that are applicable in special contexts but not in general. When they are applicable, they should be consistent with the behavioural interpretation and the theory of coherence that is developed from it.

2.3.2 Coherence on linear spaces

To introduce the notion of coherence in the simplest possible way, we now assume that the domain \mathcal{K} of \underline{P} is a linear space. Coherence can then be characterized by the three axioms in the following definition.³

2.3.3 Definition

Suppose that the domain \mathcal{K} of the lower revision \underline{P} is a linear space; that is, if $X \in \mathcal{K}$, $Y \in \mathcal{K}$ and $\lambda \in \mathbb{R}$ then $\lambda X \in \mathcal{K}$ and $X + Y \in \mathcal{K}$. Then \underline{P} is said to be **coherent** when it satisfies the three axioms:

- (P1) $\underline{P}(X) \geq \inf X$ when $X \in \mathcal{K}$ (accepting sure gains)
- (P2) $\underline{P}(\lambda X) = \lambda \underline{P}(X)$ when $X \in \mathcal{K}$ and $\lambda > 0$ (positive homogeneity)
- (P3) $\underline{P}(X + Y) \geq \underline{P}(X) + \underline{P}(Y)$ when $X \in \mathcal{K}$ and $Y \in \mathcal{K}$ (superlinearity).

Note that the definition applies when $\mathcal{K} = \mathcal{L}(\Omega)$ is the linear space of all gambles on Ω . It also applies when \underline{P} is defined for all gambles that are measurable with respect to a σ -field \mathcal{A} of subsets of Ω , since the class of \mathcal{A} -measurable gambles (section 3.2.1) is a linear space. By P2 and P3, a coherent lower revision \underline{P} is a **concave** function on its domain, i.e. $\underline{P}(\lambda X + (1 - \lambda)Y) \geq \lambda \underline{P}(X) + (1 - \lambda)\underline{P}(Y)$ when $0 \leq \lambda \leq 1$.

2.3.4 Justification of the axioms

The three coherence axioms can be justified through the behavioural interpretation of lower previsions (section 2.3.1) and the earlier axioms for desirability (section 2.2.3). Indeed, coherence axioms P1, P2, P3 correspond to the desirability axioms D1, D2, D3 respectively.⁴ Recall that $\underline{P}(X)$ is a supremum of prices You are willing to pay for X . Since paying price μ in return for X is equivalent to accepting the gamble $X - \mu$, we may say that $\underline{P}(X)$ is a supremum of prices μ for which $X - \mu$ is desirable to You.

Then P1 asserts that $X - \mu$ is desirable whenever $\inf X > \mu$, which is equivalent to $\inf(X - \mu) > 0$. In other words, any **uniform sure gain** is desirable. This follows from axiom D1.

Axiom P2 similarly follows from D2. Let \mathcal{D} denote the class of desirable gambles. When $X \in \mathcal{K}$ and $\lambda > 0$,

$$\begin{aligned}\underline{P}(\lambda X) &= \sup\{\mu: \lambda X - \mu \in \mathcal{D}\} = \sup\{\lambda\beta: \lambda X - \lambda\beta \in \mathcal{D}\} \\ &= \sup\{\lambda\beta: X - \beta \in \mathcal{D}\} \quad (\text{using D2}) \\ &= \lambda \sup\{\beta: X - \beta \in \mathcal{D}\} = \lambda \underline{P}(X).\end{aligned}$$

In effect, P2 asserts that the desirability of gambles is not affected by the utility units in which their rewards are measured.

Finally, P3 asserts that You should be willing to pay at least as much for the sum $X + Y$ as You are willing to pay for X and Y separately. This follows from D3. If $X \in \mathcal{K}$ and $Y \in \mathcal{K}$, let $\mu < \underline{P}(X)$ and $\alpha < \underline{P}(Y)$. Then $X - \mu$ and $Y - \alpha$ are each desirable by interpretation of \underline{P} , so their sum $X + Y - \mu - \alpha$ is desirable by D3, and $\underline{P}(X + Y) \geq \mu + \alpha$. Hence $\underline{P}(X + Y) \geq \underline{P}(X) + \underline{P}(Y)$.

We regard coherence as a fundamental constraint on lower previsions. Of course, it is essential to the above justification that gambles be expressed in a linear utility scale. When rewards $X(\omega)$ are amounts of money, for example, axioms D2 and P2 are not expected to hold, as attitudes to gambles may well depend on whether their units are dollars or thousands of dollars. Axioms D3 and P3 will not necessarily hold as You may be unwilling to combine several monetary transactions that are separately desirable; the argument that You should be willing to do so (used to justify D3) relied essentially on the properties of probability currency.

2.3.5 Upper previsions

Suppose that \underline{P} is a lower revision defined on a linear space \mathcal{K} . Define its **conjugate upper revision** \bar{P} on the same domain \mathcal{K} by $\bar{P}(X) = -\underline{P}(-X)$. Since $\underline{P}(Y)$ is interpreted as the supremum price μ for which $Y - \mu$ belongs

2.3 UPPER AND LOWER PREVISIONS

to the class \mathcal{D} of desirable gambles, we have

$$\begin{aligned}\bar{P}(X) &= -\sup\{\mu: -X - \mu \in \mathcal{D}\} \\ &= \inf\{-\mu: -X - \mu \in \mathcal{D}\} \\ &= \inf\{\alpha: \alpha - X \in \mathcal{D}\}.\end{aligned}$$

Thus $\bar{P}(X)$ is an infimum price α for which $\alpha - X$ is desirable, i.e. such that You are willing to sell X in return for price α . Just as the lower revision $\underline{P}(X)$ is a supremum buying price for X , the upper revision $\bar{P}(X)$ is an **infimum selling price** for X .

The preceding argument shows that selling a gamble X for price α is equivalent to buying $-X$ for price $-\alpha$. It follows that selling prices (upper revisions) are determined by buying prices (lower revisions), and vice versa, provided that $-X \in \mathcal{K}$ whenever $X \in \mathcal{K}$. The theory can therefore be developed in terms of either upper or lower previsions. We will continue to emphasize lower previsions, but use upper previsions whenever it is more convenient to do so.

When \underline{P} is coherent and \bar{P} is its conjugate, we do have

$$\inf X \leq \underline{P}(X) \leq \bar{P}(X) \leq \sup X \quad \text{for all } X \in \mathcal{K},$$

as expected. To prove that, note that the zero gamble $0 \in \mathcal{K}$, and $\underline{P}(0) = 0$ by P2. Applying P3, $\underline{P}(X + (-X)) = \underline{P}(0) = 0 \geq \underline{P}(X) + \underline{P}(-X) = \underline{P}(X) - \bar{P}(X)$. Hence $\bar{P}(X) \geq \underline{P}(X)$. By P1, $\underline{P}(-X) \geq \inf(-X) = -\sup X$, hence $\bar{P}(X) \leq \sup X$.

An upper revision \bar{P} is said to be **coherent** when its conjugate lower revision, defined by $\underline{P}(X) = -\bar{P}(-X)$, is coherent. Coherence of upper previsions is characterized by conjugate versions of axioms P1–P3:

$$(P1a) \quad \bar{P}(X) \leq \sup X \quad \text{when } X \in \mathcal{K} \quad (\text{accepting sure gains})$$

$$(P2a) \quad \bar{P}(\lambda X) = \lambda \bar{P}(X) \quad \text{when } X \in \mathcal{K} \text{ and } \lambda > 0 \quad (\text{positive homogeneity})$$

$$(P3a) \quad \bar{P}(X + Y) \leq \bar{P}(X) + \bar{P}(Y) \quad \text{when } X \in \mathcal{K} \text{ and } Y \in \mathcal{K} \quad (\text{sublinearity}).$$

Properties P2a and P3a imply that \bar{P} is a convex function on \mathcal{K} .

Economists will find it natural to regard gambles as commodities that can be bought and sold like other commodities.⁵ Our experience with other commodities is that their buying and selling prices in a market usually disagree; the commodity is sold by market traders at a higher price than they will buy it. The stock markets, insurance markets and various forms of gambling indicate that the same is true of gambles; upper previsions for gambles usually exceed lower previsions in the same market. (See section 2.9 and Appendix C for specific models.) Nevertheless, Bayesians only consider models for which buying prices agree with selling prices. In that

case the lower revision agrees with its conjugate upper revision, and it is called a **linear revision**.

2.3.6 Linear revisions

Any coherent lower revision (defined on a linear space) which is **self-conjugate** is called a linear revision. Self-conjugacy means that $\underline{P}(X) = -\underline{P}(-X)$ for all $X \in \mathcal{K}$. In that case we will drop the underbar and write $P(X)$ instead of $\underline{P}(X)$. The interpretation is that You are willing to buy X for any price less than $P(X)$, and to sell X for any price greater than $P(X)$. To use de Finetti's terminology, $P(X)$ is Your fair price for X . This definition of a linear revision is equivalent to the two axioms of de Finetti (1974, section 3.1.5):

- (i) $P(X + Y) = P(X) + P(Y)$ when $X \in \mathcal{K}$ and $Y \in \mathcal{K}$ (linearity)
- (ii) $\inf X \leq P(X) \leq \sup X$ when $X \in \mathcal{K}$ (convexity).

De Finetti proves that these two axioms imply

- (iii) $P(\lambda X) = \lambda P(X)$ when $X \in \mathcal{K}$ and $\lambda \in \mathbb{R}$.

Hence P satisfies the coherence axioms P1–P3 and is self-conjugate whenever it satisfies de Finetti's axioms. Conversely, if P is a self-conjugate coherent lower revision, (i) follows from P3 and P3a, while (ii) follows from P1 and P1a. Linear revisions are therefore a special type of lower revision. A more general definition of linear revisions will be given in section 2.8.

2.3.7 Vacuous revisions

We noted that conjugate pairs of coherent upper and lower revisions satisfy $\inf X \leq \underline{P}(X) \leq \bar{P}(X) \leq \sup X$. Linear revisions satisfy the self-conjugacy or precision condition $\underline{P}(X) = \bar{P}(X)$. At the other extreme from linearity are the **vacuous revisions**, defined by $\underline{P}(X) = \inf X$ and $\bar{P}(X) = \sup X$ for all $X \in \mathcal{L}$. Because the infimum function is positively homogeneous and superlinear, the vacuous lower revision satisfies P1–P3 and is coherent. In fact, by P1, the vacuous lower revision is the minimal coherent lower revision, and it maximizes the imprecision $\bar{P}(X) - \underline{P}(X)$ for every gamble X amongst all coherent revisions. (Compare with linear revisions, which minimize $\bar{P}(X) - \underline{P}(X)$.)

Under our behavioural interpretation, the vacuous model implies only that You are willing to accept the uniform sure gains (gambles X for which $\inf X > 0$). It makes minimal, or 'vacuous', claims about Your behavioural dispositions. It therefore seems to be the proper model for 'complete ignorance' about the true state ω .

Further examples of coherent upper and lower revisions, intermediate between these two extremes of linearity and vacuity, will be given later in this chapter, especially in section 2.9.

2.4 Avoiding sure loss

We now drop the simplifying assumption that the domain \mathcal{K} of P is a linear space and allow \mathcal{K} to be an arbitrary set of gambles. The lower revision $\underline{P}(X)$ will still be interpreted as a supremum buying price for the gamble X . The aim is to characterize coherence of the lower revision $(\Omega, \mathcal{K}, \underline{P})$ purely in terms of the values taken by \underline{P} on the domain \mathcal{K} . That will be done in section 2.5. In this section we introduce a rationality condition of 'avoiding sure loss', which is a weaker property than coherence but is of great importance in the ensuing theory. The main reason for its importance is that it is usually easier to satisfy than coherence. To make Your assessments of probabilities and revisions coherent, You must, in effect, become aware of their full implications on the domain \mathcal{K} of interest. That may be difficult to achieve directly, especially when \mathcal{K} has a complicated structure. (It should be quite easy when \mathcal{K} is a linear space.) It will often be much easier to make assessments which avoid sure loss. The theory of natural extension developed in Chapter 3 can then be used to construct a coherent model from these assessments. Provided Your assessments avoid sure loss, coherence can be achieved indirectly.

As an example of the difference between avoiding sure loss and coherence, let X_j denote the gamble which pays 1 unit if a die shows the number j , and otherwise pays nothing. Suppose You assess the lower revisions $\underline{P}(X_1) = \underline{P}(X_2) = \underline{P}(X_1 + X_2) = \frac{1}{4}$. These assessments are incoherent because they violate axiom P3. However, they avoid sure loss as defined below; the dispositions they imply cannot be combined to produce a sure loss. In this case it is easy to form the coherent natural extension of the assessments, simply by redefining $\underline{P}(X_1 + X_2) = \underline{P}(X_1) + \underline{P}(X_2) = \frac{1}{2}$. In more complicated problems it may not be so easy to detect and remove incoherence.

If You made the further assessments that $\underline{P}(X_j) = \frac{1}{4}$ for all possible values of j ($1 \leq j \leq 6$), then You would no longer avoid sure loss. You would be willing to bet at rates up to $\frac{1}{4}$ on all six possible outcomes, and the six bets taken together would produce a sure loss of up to $\frac{1}{2}$. In this case Your probability assessments concerning the die are simply irrational. Contrast this with the previous case of incoherence, where You are 'irrational' only in the much weaker sense that You have failed to recognize some of the implications of Your assessments. Incoherence reflects a kind of ignorance about the consequences of Your probability judgements, whereas failure to avoid sure loss reflects fundamental irrationality or error in judgement.¹

2.4.1 Definition of avoiding sure loss

Suppose $(\Omega, \mathcal{K}, \underline{P})$ is a lower prevision, where \mathcal{K} is an arbitrary subset of $\mathcal{L}(\Omega)$. For $X \in \mathcal{K}$, let $G(X)$ denote the gamble $X - \underline{P}(X)$. We call $G(X)$ the **marginal gamble** on X , since $G(X) + \epsilon$ is desirable to You for all positive ϵ .² (So $G(X)$ is ‘almost desirable’ – we never assume that $G(X)$ is really desirable.) Say that the lower prevision $(\Omega, \mathcal{K}, \underline{P})$ **avoids sure loss** if $\sup \sum_{j=1}^n G(X_j) \geq 0$ whenever $n \geq 1$ and the gambles X_1, X_2, \dots, X_n (not necessarily distinct) are in \mathcal{K} . Say that $(\Omega, \mathcal{K}, \underline{P})$ **incurs sure loss** if it does not avoid sure loss.

So \underline{P} incurs sure loss when $\sup \sum_{j=1}^n G(X_j) < 0$ for some X_1, \dots, X_n in \mathcal{K} . In that case, there is a positive δ such that $\sup \sum_{j=1}^n (G(X_j) + \delta) < 0$. But $G(X_j) + \delta = X_j - (\underline{P}(X_j) - \delta)$ is a desirable gamble, since it is just the gamble in which You pay $\underline{P}(X_j) - \delta$ for X_j . Thus \underline{P} incurs sure loss just when it gives rise to finitely many desirable gambles whose sum is sure to be uniformly negative. This is a ‘sure loss’, in that it is not ‘possible’ (in whatever sense of possibility is adopted) for You to avoid losing at least some positive amount of utility.³

2.4.2 Justification

The justification for avoiding sure loss is immediate from this interpretation in terms of desirability. If δ is positive, each gamble $G(X_j) + \delta$ is desirable to You, and it follows from desirability axiom D3 and induction on n that the finite sum $\sum_{j=1}^n (G(X_j) + \delta)$ is desirable. Axiom D0 then implies that $\sup \sum_{j=1}^n (G(X_j) + \delta) \geq 0$. Avoiding sure loss is therefore a consequence of axioms D0 and D3.

If Your dispositions satisfy axiom D3 and are correctly modelled by a lower prevision \underline{P} that incurs sure loss, then You are effectively disposed to give away something that has positive utility to You. That seems irrational, even if there is no practical situation in which Your ‘sure loss’ can be exploited. When rewards are expressed in a linear utility scale, we regard D3 as a necessary condition of rationality, and we therefore take the same view of avoiding sure loss.

2.4.3 Equivalent conditions

Definition 2.4.1 can be rewritten to say that \underline{P} avoids sure loss if and only if $\sup \sum_{j=1}^n X_j \geq \sum_{j=1}^n \underline{P}(X_j)$ whenever $n \geq 1$ and X_1, \dots, X_n are in \mathcal{K} . The simplest consequence of this, taking $n = 1$, is that $\sup X \geq \underline{P}(X)$ whenever $X \in \mathcal{K}$, meaning that You should not be willing to pay more for X than the supremum amount You can get back.

2.4 AVOIDING SURE LOSS

The following lemma states several conditions which are equivalent to avoiding sure loss.

2.4.4 Lemma

Suppose $(\Omega, \mathcal{K}, \underline{P})$ is a lower prevision.

- (a) \underline{P} avoids sure loss if and only if $\sup \sum_{j=1}^n \lambda_j G(X_j) \geq 0$ whenever $n \geq 1$, $X_j \in \mathcal{K}$ and $\lambda_j \geq 0$ for $1 \leq j \leq n$.⁴
- (b) Suppose $\{X(\omega) : \omega \in \Omega\}$ is finite for every $X \in \mathcal{K}$. Then \underline{P} avoids sure loss if and only if, whenever $n \geq 1$ and $X_1, \dots, X_n \in \mathcal{K}$, there exists $\omega \in \Omega$ such that $\sum_{j=1}^n G(X_j)(\omega) \geq 0$.

Proof. We prove the ‘only if’ statements; the ‘if’ parts are obvious from Definition 2.4.1.

- (a) Suppose $\sup \sum_{j=1}^n \lambda_j G(X_j) = -\delta < 0$ for some $n \geq 1$, $X_j \in \mathcal{K}$ and $\lambda_j \geq 0$. Since each X_j is bounded, there is a positive α such that $|G(X_j)(\omega)| \leq \alpha$ for all $\omega \in \Omega$ and $1 \leq j \leq n$. Let $\epsilon = \delta/2n\alpha$. For each j , find some rational ρ_j such that $\lambda_j \leq \rho_j \leq \lambda_j + \epsilon$. Then

$$\sum_{j=1}^n \rho_j G(X_j) \leq \sum_{j=1}^n (\lambda_j G(X_j) + \epsilon\alpha) \leq -\delta + n\epsilon\alpha = -\delta/2.$$

Now the ρ_j are non-negative rationals, and we can multiply through by a common denominator k to give $\sup \sum_{j=1}^n m_j G(X_j) \leq -k\delta/2$, where now m_j are non-negative integers. Thus \underline{P} incurs sure loss.

- (b) Suppose there are $n \geq 1, X_1, \dots, X_n \in \mathcal{K}$ with $Y(\omega) = \sum_{j=1}^n G(X_j)(\omega) < 0$ for all $\omega \in \Omega$. Since each X_j takes only finitely many values, so does $G(X_j)$ and so does Y . Hence $\sup Y < 0$, so that \underline{P} incurs sure loss. ♦

The non-negative coefficients λ_j in (a) of the lemma play the role of **stakes** on the marginal gambles $G(X_j)$. In condition (a) the gambles X_j can obviously be assumed to be distinct, whereas that cannot be assumed in the original definition of avoiding sure loss. On the other hand, the original definition has the advantage of making it clear that any sure loss can be exploited through gambles with integer stakes. That is especially important when all rewards and prices are restricted to be integer multiples of some basic unit of currency (e.g. lottery tickets or dollars); then the gambles making up a sure loss can be assumed to involve only integer multiples of the same unit.

Part (b) of the lemma shows that $\sup \sum_{j=1}^n G(X_j) \geq 0$ in the definition of avoiding sure loss can be replaced by $\sum_{j=1}^n G(X_j)(\omega) \geq 0$ for some $\omega \in \Omega$, provided that each gamble in \mathcal{K} takes only finitely many distinct values. That is so, for example, when \mathcal{K} consists of indicator functions of events. The following example shows that the proviso is needed. When some gamble

in \mathcal{K} takes infinitely many distinct values, \underline{P} may ‘avoid sure loss’ even though $\sum_{j=1}^n G(X_j)(\omega) < 0$ for every ω in Ω . This indicates that avoiding sure loss is a rather weak requirement which might, more accurately, be termed ‘avoiding uniform sure loss’.⁵

2.4.5 Example

Let $\Omega = \mathbb{Z}^+$ (the set of positive integers), define X by $X(\omega) = -1/\omega$ for $\omega \in \Omega$, let $\mathcal{K} = \{X\}$ and $\underline{P}(X) = 0$. The marginal gamble on X is $G(X) = X$. Then $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss, according to Definition 2.4.1, because $\sup G(X) = \sup X = 0$. (Indeed \underline{P} is coherent.) But $X(\omega) < 0$ for all $\omega \in \Omega$, so You are sure to lose from accepting the marginal gamble $G(X)$.

Readers may feel that the assessment $\underline{P}(X) = 0$ here is irrational, and that the definition of avoiding sure loss (section 2.4.1) should be strengthened to rule it out. Stronger properties of strict coherence or countable coherence could be introduced, and these would force $\underline{P}(X) < 0$ in this example. However, these conditions are less convincing than Definition 2.4.1.⁶ Nor is it obvious that there is anything ‘irrational’ about the assessment $\underline{P}(X) = 0$. Under the interpretation we have given, this assessment expresses Your willingness to accept any gamble $X + \delta$ when δ is positive. You need not be willing to accept X itself, and of course You would not be, since You can only lose by doing so. Now $X(\omega) + \delta > 0$ when $\omega > 1/\delta$, so You might judge the gamble $X + \delta$ to be desirable if You believed the integer ω was probably larger than $1/\delta$. If You believed this for all positive δ then clearly Your beliefs could not be modelled by a countably additive probability measure, but they could be modelled by finitely additive distributions which assign zero probability to every finite set of integers. (See the model for a ‘uniform distribution’ on the positive integers, in section 2.9.5.) At this point we see no compelling reason to rule out probability models that are not countably additive. (However, more convincing reasons for doing so will appear when we examine conditional previsions, in Chapter 6.)

2.4.6 Upper previsions

If $(\Omega, \mathcal{K}, \underline{P})$ is a lower prevision, the **conjugate upper prevision** \bar{P} is defined on the domain $-\mathcal{K} = \{X : -X \in \mathcal{K}\}$ by $\bar{P}(X) = -\underline{P}(-X)$. Note that \bar{P} may be defined on a different domain from \underline{P} . Whenever \underline{P} avoids sure loss and both X and $-X$ are in \mathcal{K} , we have (taking $n = 2$, $X_1 = X$, $X_2 = -X$ in Definition 2.4.1) $\underline{P}(X) + \underline{P}(-X) \leq 0$, so that $\underline{P}(X) \leq \bar{P}(X)$. So a lower prevision which avoids sure loss is never larger than its conjugate upper prevision.

For $X \in -\mathcal{K}$ we can write $G(-X) = -X - \underline{P}(-X) = \bar{P}(X) - X$. The marginal gamble on $-X$ can therefore be regarded as a transaction in which You are paid a price $\bar{P}(X)$ in return for the gamble X . The definition of avoiding sure loss can be written equivalently in terms of \bar{P} . It requires that $\sup \sum_{j=1}^n [\bar{P}(X_j) - X_j] \geq 0$ whenever $n \geq 1$ and X_1, \dots, X_n are in $-\mathcal{K}$.

2.4.7 Consequences of avoiding sure loss

Suppose the lower prevision $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss and \bar{P} is its conjugate upper prevision. The following properties hold whenever their terms are well defined (i.e., all gambles involved are in the domains of \underline{P} or \bar{P}):

- (a) $\underline{P}(X) \leq \sup X$, $\bar{P}(X) \geq \inf X$, $\underline{P}(X) \leq \bar{P}(X)$
- (b) $(\forall \mu \in \mathbb{R}) \quad \underline{P}(\mu) \leq \mu \leq \bar{P}(\mu)$
- (c) $(\forall \mu \in \mathbb{R}) \quad \underline{P}(X) + \underline{P}(\mu - X) \leq \mu$
- (d) $(\forall \mu \in \mathbb{R}) \quad$ if $X \geq Y + \mu$ then $\bar{P}(X) \geq \underline{P}(Y) + \mu$
- (e) $\underline{P}(X + Y) \leq \bar{P}(X) + \bar{P}(Y)$, $\bar{P}(X + Y) \geq \underline{P}(X) + \underline{P}(Y)$
- (f) $(\forall \lambda \geq 0) \quad \underline{P}(\lambda X) \leq \lambda \bar{P}(X)$, $\bar{P}(\lambda X) \geq \lambda \underline{P}(X)$
- (g) $(\forall 0 \leq \lambda \leq 1) \quad \underline{P}(\lambda X + (1 - \lambda)Y) \leq \lambda \bar{P}(X) + (1 - \lambda) \bar{P}(Y)$,
 $\bar{P}(\lambda X + (1 - \lambda)Y) \geq \lambda \underline{P}(X) + (1 - \lambda) \underline{P}(Y)$
- (h) $(\forall \lambda_1, \dots, \lambda_n \geq 0) \quad \underline{P}\left(\sum_{j=1}^n \lambda_j X_j\right) \leq \sum_{j=1}^n \lambda_j \bar{P}(X_j)$,
 $\bar{P}\left(\sum_{j=1}^n \lambda_j X_j\right) \geq \sum_{j=1}^n \lambda_j \underline{P}(X_j)$
- (i) $(\forall \lambda_1, \dots, \lambda_n \geq 0 \text{ and } \mu \in \mathbb{R})$
if $X_0 \geq \sum_{j=1}^n \lambda_j X_j + \mu$ then $\bar{P}(X_0) \geq \sum_{j=1}^n \lambda_j \bar{P}(X_j) + \mu$,
if $X_0 \leq \sum_{j=1}^n \lambda_j X_j + \mu$ then $\underline{P}(X_0) \leq \sum_{j=1}^n \lambda_j \underline{P}(X_j) + \mu$
- (j) if $\lambda_1, \dots, \lambda_n, \beta_1, \dots, \beta_m \geq 0$ and $\sum_{j=1}^n \lambda_j X_j \leq \sum_{i=1}^m \beta_i Y_i$ then
 $\sum_{j=1}^n \lambda_j \bar{P}(X_j) \leq \sum_{i=1}^m \beta_i \bar{P}(Y_i)$.

Proof. Apply Definition 2.4.1 with (a) $X_1 = X$, then $X_1 = X$, $X_2 = -X$; (c) $X_1 = X$, $X_2 = \mu - X$; (d) $X_1 = -X$, $X_2 = Y$; (e) $X_1 = -X$, $X_2 = -Y$, $X_3 = X + Y$. (b) follows from (a), and the other parts of (a)–(e) follow from the conjugacy relation. For (f)–(j), apply condition 2.4.4(a) with (f) $X_1 = \lambda X$, $X_2 = -X$, $\lambda_1 = 1$, $\lambda_2 = \lambda$; (g) $X_1 = \lambda X + (1 - \lambda)Y$, $X_2 = -X$, $X_3 = -Y$,

$\lambda_1 = 1, \lambda_2 = \lambda, \lambda_3 = 1 - \lambda$; (h) $X_{n+1} = -\sum_{j=1}^n \lambda_j X_j, \lambda_{n+1} = 1$; (i) $X_{n+1} = -X_0, \lambda_{n+1} = 1$; (j) $X_{n+i} = -Y_i$ and $\lambda_{n+i} = \beta_i$ for $1 \leq i \leq m$; and use the conjugacy relation. ♦

2.4.8 Arbitrage

Simple examples of prices that ‘incur sure loss’ can be found in the financial markets. Arbitrageurs earn their living by recognizing these situations and exploiting them to make a sure gain. In **classical arbitrage**, which was apparently first practised by European currency traders in the late Middle Ages, commodities, securities or foreign exchange are bought in one market and almost simultaneously sold in another market at a higher price. More sophisticated kinds of arbitrage have recently become popular on the stock markets. These involve the simultaneous buying and selling of various risky investments, such as stocks, convertible bonds, put or call options, and stock or commodity futures contracts, in such a way as to guarantee a sure gain (irrespective of changes in the values of the individual investments).⁷

For example, **stock-index-futures arbitrage** is possible when there is a discrepancy between a current stock index (which is just a weighted average of many individual stock prices) and the price of the corresponding ‘stock index future’, which is effectively a gamble X whose reward is the level of the stock index at a specified future date. Suppose the current stock index, α , is less than the current price of the stock index future, μ . An arbitrageur can buy the individual stocks that make up the stock index for price α , and simultaneously sell the stock index future for price μ .⁸ At the future date, the profit from buying the stocks is $X - \alpha$, and the profit from selling the stock index future is $\mu - X$. Each investment is risky, but together they yield a sure gain of $\mu - \alpha$ for the arbitrageur. Here μ is a market buying price for X and α is effectively a market selling price for X . The market prices incur sure loss because $\underline{P}(X) = \mu > \alpha = \bar{P}(X)$.

2.5 Coherence

Next we strengthen the definition of avoiding sure loss by giving a general definition of coherence. This will be shown to agree with the earlier definition (2.3.3) in the special case where the domain is a linear space.

To motivate the definition, consider the simple example of incoherence given at the start of section 2.4, in which $\underline{P}(X_1) = \underline{P}(X_2) = \underline{P}(X_1 + X_2) = \frac{1}{4}$. According to this model, You are willing to pay up to $\frac{1}{4}$ for each of X_1 and X_2 . Hence You should be willing to pay up to $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ for $X_1 + X_2$. Because this is more than the supremum buying price $\underline{P}(X_1 + X_2) = \frac{1}{4}$, the model is incoherent. The meaning of incoherence is that the specified buying

prices $\underline{P}(X_j)$ effectively imply a buying price for some other gamble X_0 that is higher than the specified price $\underline{P}(X_0)$.

We can rephrase the preceding argument in terms of the marginal gambles $G(X_1), G(X_2)$ and $G(X_1 + X_2)$. Here $G(X_1) + G(X_2) = X_1 + X_2 - \frac{1}{2} = G(X_1 + X_2) - \frac{1}{4}$. By D3, the gamble $G(X_1) + G(X_2)$ should be (at least) marginally desirable, hence so should the equivalent gamble $G(X_1 + X_2) - \frac{1}{4}$. This is incoherent because $G(X_1 + X_2)$ was asserted to be only marginally desirable.

The crux of this argument is that $G(X_1 + X_2) - \delta \geq G(X_1) + G(X_2)$ for some positive δ . More generally, incoherence exists whenever there are gambles X_0, X_1, \dots, X_n and a positive δ such that $G(X_0) - \delta \geq \sum_{j=1}^n G(X_j)$. A slight extension of this leads to the following definition.

2.5.1 General definition of coherence¹

Let \mathcal{K} be an arbitrary subset of $\mathcal{L}(\Omega)$. Say that a lower prevision $(\Omega, \mathcal{K}, \underline{P})$ is **coherent** if $\sup[\sum_{j=1}^n G(X_j) - mG(X_0)] \geq 0$ whenever m and n are non-negative integers and X_0, X_1, \dots, X_n (not necessarily distinct) are in \mathcal{K} .

2.5.2 Justification

We wish to justify coherence as a rationality requirement, using axioms D1–D3 of section 2.2.3. To do so we break down the condition into three cases.

Case 1: $m = 0$

In this case coherence reduces to avoiding sure loss (Definition 2.4.1), which was justified in section 2.4.2. Thus every coherent lower prevision avoids sure loss.²

Case 2: $m > 0$ but $n = 0$

Coherence reduces to $\sup[-G(X_0)] = -\inf G(X_0) \geq 0$, i.e., $\underline{P}(X_0) \geq \inf X_0$ for every $X_0 \in \mathcal{K}$. This simply means that all the gambles $X_0 - \mu$ with $\mu < \inf X_0$, which are uniform sure gains, are desirable to You. It follows from axiom D1.

Case 3: $m > 0$ and $n > 0$ ³

To justify the condition in this case, consider what happens when it fails. Then there are positive integers m and n , gambles X_0, X_1, \dots, X_n in \mathcal{K} , and some positive δ such that

$$\sum_{j=1}^n [G(X_j) + \delta] \leq m[G(X_0) - \delta].$$

The gamble on the left-hand side of this inequality should be desirable to You, by axiom D3, since each $G(X_j) + \delta$ is desirable. The better gamble $m[G(X_0) - \delta]$ should therefore be desirable. Using axiom D2, the positive multiple $G(X_0) - \delta$ should be desirable. But $G(X_0) - \delta = X_0 - (\underline{P}(X_0) + \delta)$, so this means that You should be disposed (in view of Your other dispositions) to pay at least the price $\underline{P}(X_0) + \delta$ for X_0 . This price is higher than the supremum buying price $\underline{P}(X_0)$ specified by the model. Thus the model $(\Omega, \mathcal{K}, \underline{P})$ is inconsistent. The coherence condition 2.5.1 rules out this type of inconsistency.

The argument given in case 3 is constructive, in that it shows us how to correct an incoherent lower prevision. Whenever coherence fails for a gamble X_0 , we can replace the given supremum buying price $\underline{P}(X_0)$ by some higher buying price which can be constructed by considering all gambles of the form $\sum_{j=1}^n G(X_j) - mX_0$. Provided the given lower prevision \underline{P} avoids sure loss, the corrected lower prevision constructed from it, which is called the **natural extension** of \underline{P} , will be coherent. The natural extension summarizes the buying prices that are implied by the specified prices \underline{P} .

It should now be clear that coherence characterizes a kind of self-consistency of a lower prevision. Like avoiding sure loss, coherence is proposed as a normative criterion of rationality. It is, however, a much stronger condition than avoiding sure loss. To achieve coherence You must work out the full implications for buying prices of Your initial assessments \underline{P} . That can be done, provided \underline{P} avoids sure loss, by natural extension of \underline{P} . This should allay any suspicions that coherence, in demanding rather complete knowledge about the implications of specified previsions, is too strong a requirement. In practice You will be required only to satisfy the weaker condition of avoiding sure loss, and the mathematical theory of natural extension can be used to achieve coherence.

2.5.3 Equivalent conditions

By substituting for the marginal gambles $G(X_j) = X_j - \underline{P}(X_j)$, the coherence condition may be written as:

$$\sup \left[\sum_{j=1}^n X_j - mX_0 \right] \geq \sum_{j=1}^n \underline{P}(X_j) - m\underline{P}(X_0)$$

whenever m and n are non-negative integers and $X_0, X_1, \dots, X_n \in \mathcal{K}$.

Two further conditions equivalent to coherence are stated in the following lemma. Compare these with the corresponding conditions for avoiding sure loss in Lemma 2.4.4. (The proof is a simple extension of 2.4.4 and is omitted.)

2.5.4 Lemma

Suppose $(\Omega, \mathcal{K}, \underline{P})$ is a lower prevision.

- (a) \underline{P} is coherent if and only if $\sup[\sum_{j=1}^n \lambda_j G(X_j) - \lambda_0 G(X_0)] \geq 0$ whenever $n \geq 1$, $X_j \in \mathcal{K}$ and $\lambda_j \geq 0$ for $0 \leq j \leq n$.
- (b) Suppose $\{X(\omega) : \omega \in \Omega\}$ is finite for every $X \in \mathcal{K}$. Then \underline{P} is coherent if and only if, whenever m and n are non-negative integers and $X_0, X_1, \dots, X_n \in \mathcal{K}$, there exists $\omega \in \Omega$ such that $\sum_{j=1}^n G(X_j)(\omega) \geq mG(X_0)(\omega)$.

Condition (a) can be used to check for coherence of a lower prevision \underline{P} that is specified numerically on finite Ω and \mathcal{K} . An algorithm for this purpose is outlined in Appendix A. It involves solving sets of n linear equations in n or fewer unknowns, where n is the cardinality of Ω .

As usual, all these conditions can be written in terms of the conjugate upper prevision $\bar{P}(X) = -\underline{P}(-X)$. The definition of coherence becomes, in terms of \bar{P} : $\inf[\sum_{j=1}^n X_j - mX_0] \leq \sum_{j=1}^n \bar{P}(X_j) - m\bar{P}(X_0)$ whenever m and n are non-negative integers and each $X_j \in -\mathcal{K}$.

We show next that the general definition of coherence (2.5.1) is equivalent to the earlier definition (2.3.3) when the domain \mathcal{K} is a linear space. That should be plausible to the reader, as exactly the same axioms (D1–D3) were needed to justify each definition.

2.5.5 Theorem

Suppose \mathcal{K} is a linear subspace of $\mathcal{L}(\Omega)$. Then $(\Omega, \mathcal{K}, \underline{P})$ is coherent, in the sense of the general definition 2.5.1, if and only if it satisfies axioms P1, P2 and P3.

Proof. Suppose first that \underline{P} satisfies condition 2.5.1. To verify P1, take $n = 0$, $m = 1$ and $X_0 = X$ in 2.5.1. For P3, take $n = 2$, $m = 1$, $X_0 = X + Y$, $X_1 = X$, $X_2 = Y$ in 2.5.1. For P2, let $n = 1$, $X_0 = X$, $X_1 = \lambda X$, $\lambda_0 = \lambda$, $\lambda_1 = 1$ in condition 2.5.4(a), to show $\lambda \underline{P}(X) \geq \underline{P}(\lambda X)$. Then interchange (X_0, λ_0) with (X_1, λ_1) to get $\underline{P}(\lambda X) \geq \lambda \underline{P}(X)$.

Conversely, suppose that \underline{P} satisfies the three axioms. Let m and n be non-negative integers, and let $X_0, X_1, \dots, X_n \in \mathcal{K}$. We show that condition 2.5.1 holds. Write $X = mX_0$, $Y = \sum_{j=1}^n X_j$ (taking X or Y to be zero when m or n is zero) and $Z = X - Y$. Then X, Y, Z are in \mathcal{K} since \mathcal{K} is a linear space. Also $\underline{P}(X) = m\underline{P}(X_0)$ by P2 (or from $\underline{P}(0) = 0$ if $m = 0$), $\underline{P}(Y) \geq \sum_{j=1}^n \underline{P}(X_j)$ by P3, $\underline{P}(X) = \underline{P}(Y + Z) \geq \underline{P}(Y) + \underline{P}(Z)$ by P3, and $\underline{P}(Z) \geq \inf Z$ by P1. Hence $\sup[\sum_{j=1}^n X_j - mX_0] = \sup(Y - X) = \sup(-Z) = -\inf Z \geq -\underline{P}(Z) \geq \underline{P}(Y) - \underline{P}(X) \geq \sum_{j=1}^n \underline{P}(X_j) - m\underline{P}(X_0)$. Hence $\sup[\sum_{j=1}^n G(X_j) - mG(X_0)] \geq 0$, so 2.5.1 holds. ♦

The general definition of coherence simplifies also under weaker restrictions on the domain \mathcal{K} . For example, call \mathcal{K} a **convex cone** if $\lambda X \in \mathcal{K}$ and $X + Y \in \mathcal{K}$ whenever $\lambda \geq 0$, $X \in \mathcal{K}$ and $Y \in \mathcal{K}$. A convex cone need not satisfy the extra property $\mathcal{K} = -\mathcal{K}$ of a linear space. The set of all non-negative gambles is a convex cone but not a linear space. When \mathcal{K} is a convex cone, the lower prevision $(\Omega, \mathcal{K}, \underline{P})$ is coherent if and only if it satisfies axioms P2, P3 and the monotonicity axiom:⁴

$$(P4) \text{ if } X \in \mathcal{K}, Y \in \mathcal{K}, \mu \in \mathbb{R} \text{ and } X \geq Y + \mu \text{ then } \underline{P}(X) \geq \underline{P}(Y) + \mu.$$

Under the still weaker assumptions that \mathcal{K} is closed under addition and contains the zero gamble, coherence is equivalent to the three axioms P3, P4 and the weakened version of P2:

$$(P2') \underline{P}(mX) = m\underline{P}(X) \text{ whenever } X \in \mathcal{K} \text{ and } m \in \mathbb{Z}^+, \text{ where } mX \text{ is defined inductively by } 1X = X \text{ and } (m+1)X = mX + X. \text{⁵}$$

2.6 Basic properties of coherent previsions

The following properties of coherent upper and lower previsions will be used throughout the book, especially the basic properties (a)–(f). Compare these with the weaker consequences of avoiding sure loss (2.4.7).

2.6.1 Consequences of coherence

Suppose $(\Omega, \mathcal{K}, \underline{P})$ is a coherent lower prevision. Let \bar{P} be the conjugate upper prevision, defined on $-\mathcal{K}$ by $\bar{P}(X) = -\underline{P}(-X)$. The following properties hold whenever all gambles involved are in the domains of \underline{P} or \bar{P} :

- (a) $\inf X \leq \underline{P}(X) \leq \bar{P}(X) \leq \sup X$
- (b) $(\forall \mu \in \mathbb{R}) \underline{P}(\mu) = \bar{P}(\mu) = \mu$
- (c) $(\forall \mu \in \mathbb{R}) \underline{P}(X + \mu) = \underline{P}(X) + \mu, \bar{P}(X + \mu) = \bar{P}(X) + \mu$
- (d) $(\forall \mu \in \mathbb{R})$ if $X \geq Y + \mu$ then $\underline{P}(X) \geq \underline{P}(Y) + \mu$ and $\bar{P}(X) \geq \bar{P}(Y) + \mu$
- (e) $\underline{P}(X) + \underline{P}(Y) \leq \underline{P}(X + Y) \leq \underline{P}(X) + \bar{P}(Y) \leq \bar{P}(X + Y) \leq \bar{P}(X) + \bar{P}(Y)$
- (f) $(\forall \lambda \geq 0) \underline{P}(\lambda X) = \lambda \underline{P}(X), \bar{P}(\lambda X) = \lambda \bar{P}(X)$
- (g) $(\forall 0 \leq \lambda \leq 1) \lambda \underline{P}(X) + (1 - \lambda) \underline{P}(Y) \leq \underline{P}(\lambda X + (1 - \lambda) Y) \leq \lambda \bar{P}(X) + (1 - \lambda) \bar{P}(Y)$
- (h) $\underline{P}(X) \leq \underline{P}(|X|), \bar{P}(X) \leq \bar{P}(|X|)$, where $|X|(\omega) = |X(\omega)|$
- (i) $|\underline{P}(X) - \underline{P}(Y)| \leq \bar{P}(|X - Y|), |\bar{P}(X) - \bar{P}(Y)| \leq \bar{P}(|X - Y|)$
- (j) $\underline{P}(|X + Y|) \leq \underline{P}(|X|) + \bar{P}(|Y|), \bar{P}(|X + Y|) \leq \bar{P}(|X|) + \bar{P}(|Y|)$
- (k) Define $X \vee Y = \max\{X, Y\}$ and $X \wedge Y = \min\{X, Y\}$ by

2.6 BASIC PROPERTIES OF COHERENT PREVISIONS

$(X \vee Y)(\omega) = \max\{X(\omega), Y(\omega)\}, (X \wedge Y)(\omega) = \min\{X(\omega), Y(\omega)\}$. Then

$$\underline{P}(X \vee Y) + \underline{P}(X \wedge Y) \leq \underline{P}(X) + \bar{P}(Y) \leq \bar{P}(X \vee Y) + \bar{P}(X \wedge Y),$$

$$\underline{P}(X) + \underline{P}(Y) \leq \underline{P}(X \vee Y) + \bar{P}(X \wedge Y) \leq \bar{P}(X) + \bar{P}(Y),$$

$$\underline{P}(X) + \underline{P}(Y) \leq \bar{P}(X \vee Y) + \underline{P}(X \wedge Y) \leq \bar{P}(X) + \bar{P}(Y)$$

(l) If $\bar{P}(|X_n - X|) \rightarrow 0$ as $n \rightarrow \infty$ then $\underline{P}(X_n) \rightarrow \underline{P}(X)$ and $\bar{P}(X_n) \rightarrow \bar{P}(X)$.

(This holds, for example, if $\sup |X_n - X| \rightarrow 0$.)

$$(m) (\forall \lambda_1, \dots, \lambda_n \geq 0) \quad \underline{P}\left(\sum_{j=1}^n \lambda_j X_j\right) \geq \sum_{j=1}^n \lambda_j \underline{P}(X_j),$$

$$\text{and} \quad \bar{P}\left(\sum_{j=1}^n \lambda_j X_j\right) \leq \sum_{j=1}^n \lambda_j \bar{P}(X_j)$$

$$(n) (\forall \lambda_1, \dots, \lambda_n \geq 0 \text{ and } \mu \in \mathbb{R}) \text{ if } X_0 \geq \sum_{j=1}^n \lambda_j X_j + \mu \text{ then } \underline{P}(X_0) \geq$$

$$\sum_{j=1}^n \lambda_j \underline{P}(X_j) + \mu; \text{ if } X_0 \leq \sum_{j=1}^n \lambda_j X_j + \mu \text{ then } \bar{P}(X_0) \leq \sum_{j=1}^n \lambda_j \bar{P}(X_j) + \mu$$

$$(o) \text{ If } \lambda_1, \dots, \lambda_n, \beta_1, \dots, \beta_m \geq 0 \text{ and } \sum_{j=1}^n \lambda_j X_j \leq \sum_{i=1}^m \beta_i Y_i \text{ then } \sum_{j=1}^n \lambda_j \underline{P}(X_j) \leq \beta_1 \underline{P}(Y_1) + \sum_{i=2}^m \beta_i \bar{P}(Y_i), \text{ and } \lambda_1 \bar{P}(X_1) + \sum_{j=2}^n \lambda_j \underline{P}(X_j) \leq \sum_{i=1}^m \beta_i \bar{P}(Y_i).$$

Proof. We prove only the conditions involving \underline{P} ; the other conditions then follow from the conjugacy property $\bar{P}(X) = -\underline{P}(-X)$. To prove (a)–(e), use Definition 2.5.1 with (a) $n = 0, m = 1, X_0 = X$; (d) $n = m = 1, X_0 = X, X_1 = Y$; (e) $n = 2, m = 1, X_1 = X, X_2 = Y, X_0 = X + Y$ (the other inequalities in (e) then follow by conjugacy); (b) follows from (a); (c) follows from (d). For (f), see the proof of Theorem 2.5.5. (g) follows from (e) and (f). For (h), use $|X| \geq X$ and (d). For (i), use (e) and (h) to show $|\underline{P}(X) - \underline{P}(Y)| = \max\{\underline{P}(X) - \underline{P}(Y), \underline{P}(Y) - \underline{P}(X)\} \leq \max\{\bar{P}(X - Y), \bar{P}(Y - X)\} \leq \bar{P}(|X - Y|)$. (j) follows from (d) and (e) since $|X + Y| \leq |X| + |Y|$. For (k), use 2.5.1 with $n = 3, m = 1, X_0 = X, X_1 = X \vee Y, X_2 = X \wedge Y, X_3 = -Y$ (or similar choices), noting that $(X \vee Y) + (X \wedge Y) = X + Y$. (l) follows from (i). (m) follows from (f), (e) and induction on n . Property (n) is equivalent to coherence by Lemma 2.5.4(a), and (o) follows from 2.5.4(a) by taking $X_0 = Y_1$ or $X_0 = -X_1$. ♦

Next we describe three general methods for constructing new coherent previsions from old ones: as lower envelopes, convex combinations, or limits of sequences. Note also that the restriction of a coherent prevision to a smaller domain is always coherent. (That is clear from the definition of

coherence.) In Chapter 3 we show that any coherent prevision has a coherent extension to any larger domain.

The first result is that the lower envelope of any collection of coherent lower previsions is itself a coherent lower predition. First we define the notions of dominance and lower envelope, which have an important role in the ensuing theory.

2.6.2 Definitions

If P and Q are real-valued functions defined on domain \mathcal{K} , say that P **dominates** Q (on \mathcal{K}) when $P(X) \geq Q(X)$ for all $X \in \mathcal{K}$. In that case, write $P \geq Q$. If $\{\underline{P}_\gamma; \gamma \in \Gamma\}$ is any non-empty collection of lower previsions on domain \mathcal{K} , its **lower envelope** \underline{P} is defined on domain \mathcal{K} by $\underline{P}(X) = \inf\{\underline{P}_\gamma(X); \gamma \in \Gamma\}$.¹

Consider lower previsions $\underline{P}, \underline{Q}, \underline{P}_\gamma$ defined on \mathcal{K} , and let $\bar{P}, \bar{Q}, \bar{P}_\gamma$ denote the conjugate upper previsions on $-\mathcal{K}$. Then clearly $\underline{P} \geq \underline{Q}$ (on \mathcal{K}) if and only if $\bar{P} \leq \bar{Q}$ (on $-\mathcal{K}$). When \underline{P} dominates \underline{Q} , all the behavioural dispositions implied by \underline{Q} are also implied by \underline{P} , and we say that \underline{P} is a **more precise model than \underline{Q}** . If $\mathcal{K} = -\mathcal{K}$, the intervals $[\underline{P}(X), \bar{P}(X)]$ are contained in the corresponding intervals $[\underline{Q}(X), \bar{Q}(X)]$ for all $X \in \mathcal{K}$.

Similarly, \bar{P} is the lower envelope of $\{\bar{P}_\gamma; \gamma \in \Gamma\}$ if and only if \bar{P} is the **upper envelope** of $\{\bar{P}_\gamma; \gamma \in \Gamma\}$, meaning $\bar{P}(X) = \sup\{\bar{P}_\gamma(X); \gamma \in \Gamma\}$ for all $X \in -\mathcal{K}$.

2.6.3 Lower envelope theorem

- (a) If \underline{P} and \underline{Q} are lower previsions with domain \mathcal{K} , \underline{P} dominates \underline{Q} on \mathcal{K} and \underline{P} avoids sure loss, then \underline{Q} avoids sure loss.
- (b) Let Γ be an arbitrary non-empty index set. If \underline{P}_γ is a coherent lower predition with domain \mathcal{K} for every $\gamma \in \Gamma$, and \underline{P} is the lower envelope of $\{\underline{P}_\gamma; \gamma \in \Gamma\}$, then \underline{P} is a coherent lower predition on \mathcal{K} .

Proof. (a) Let $X_1, \dots, X_n \in \mathcal{K}$. Then $\sum_{j=1}^n \underline{Q}(X_j) \leq \sum_{j=1}^n \underline{P}(X_j) \leq \sup \sum_{j=1}^n X_j$, since \underline{P} avoids sure loss. Thus \underline{Q} avoids sure loss.

(b) Note first that each $\underline{P}_\gamma(X) \geq \inf X$, so $\underline{P}(X)$ is finite. Given non-negative integers m and n , and $X_0, X_1, \dots, X_n \in \mathcal{K}$, we need to show that $\sup [\sum_{j=1}^n X_j - mX_0] \geq \sum_{j=1}^n \underline{P}(X_j) - m\underline{P}(X_0)$. The case $m = 0$ is covered by (a), since each \underline{P}_γ dominates \underline{P} and avoids sure loss. If $m \geq 1$ and $\delta > 0$, there is $\gamma \in \Gamma$ such that $\underline{P}_\gamma(X_0) \leq \underline{P}(X_0) + \delta/m$. Then $\sum_{j=1}^n \underline{P}(X_j) - m\underline{P}(X_0) - \delta \leq \sum_{j=1}^n \underline{P}(X_j) - m\underline{P}_\gamma(X_0) \leq \sum_{j=1}^n \underline{P}_\gamma(X_j) - m\underline{P}_\gamma(X_0) \leq \sup [\sum_{j=1}^n X_j - mX_0]$ because \underline{P}_γ is coherent. Since δ can be arbitrarily small, this establishes coherence of \underline{P} . ♦

2.6 BASIC PROPERTIES OF COHERENT PREVISIONS

The lower envelope theorem is useful both for constructing coherent previsions and for verifying that given previsions are coherent or avoid sure loss. In particular, we can prove that a given lower predition avoids sure loss by showing that it is dominated by some linear predition, and we can prove that it is coherent by showing that it is a lower envelope of some class of linear previsions.

The next result asserts that the class of all coherent lower previsions on domain \mathcal{K} is convex.

2.6.4 Convexity theorem²

If \underline{P}_1 and \underline{P}_2 are lower previsions on the same domain \mathcal{K} and $0 < \lambda < 1$, define their convex combination \underline{P} on \mathcal{K} by $\underline{P}(X) = \lambda \underline{P}_1(X) + (1 - \lambda) \underline{P}_2(X)$. If \underline{P}_1 and \underline{P}_2 each avoid sure loss then so does \underline{P} . If \underline{P}_1 and \underline{P}_2 are each coherent then so is \underline{P} .

Proof. Suppose $\underline{P}_1, \underline{P}_2$ each avoid sure loss. Let $X_1, \dots, X_n \in \mathcal{K}$. Therefore $\sup \sum_{j=1}^n X_j \geq \max \{\sum_{j=1}^n \underline{P}_1(X_j), \sum_{j=1}^n \underline{P}_2(X_j)\} \geq \lambda \sum_{j=1}^n \underline{P}_1(X_j) + (1 - \lambda) \sum_{j=1}^n \underline{P}_2(X_j) = \sum_{j=1}^n \underline{P}(X_j)$. Thus \underline{P} avoids sure loss. The proof of coherence is similar. ♦

For example, any convex combination of a linear predition with the vacuous predition is coherent (see section 2.9.2). If \underline{P} is a convex combination of \underline{P}_1 and \underline{P}_2 , its conjugate \bar{P} is the same convex combination of the conjugates \bar{P}_1 and \bar{P}_2 .

Finally, we consider sequences of lower previsions.

2.6.5 Convergence theorem³

Let $\{\underline{P}_i; i \geq 1\}$ be a sequence of lower previsions (each defined on \mathcal{K}) which converges pointwise to a lower predition \underline{P} on \mathcal{K} , meaning that the sequence of real numbers $\{\underline{P}_i(X); i \geq 1\}$ converges to $\underline{P}(X)$ for every $X \in \mathcal{K}$.

- (a) If \underline{P}_i avoids sure loss for all $i \geq 1$ then \underline{P} avoids sure loss.
- (b) If \underline{P}_i is coherent for all $i \geq 1$ then \underline{P} is coherent.

Proof. (a) Given $X_1, \dots, X_n \in \mathcal{K}$ and $\delta > 0$, there is a sufficiently large i such that $|\underline{P}_i(X_j) - \underline{P}(X_j)| \leq \delta/n$ for $1 \leq j \leq n$. Then $\sup \sum_{j=1}^n X_j \geq \sum_{j=1}^n \underline{P}_i(X_j) \geq \sum_{j=1}^n \underline{P}(X_j) - \delta$. Hence \underline{P} avoids sure loss.

(b) Given non-negative integers m and n , $X_0, X_1, \dots, X_n \in \mathcal{K}$, and $\delta > 0$, there is i such that $|\underline{P}_i(X_j) - \underline{P}(X_j)| \leq \delta/(m+n)$ for $0 \leq j \leq n$. Then $\sup [\sum_{j=1}^n X_j - mX_0] \geq \sum_{j=1}^n \underline{P}_i(X_j) - m\underline{P}_i(X_0) \geq \sum_{j=1}^n \underline{P}(X_j) - m\underline{P}(X_0) - \delta$. Hence \underline{P} is coherent. ♦

2.6.6 Corollary

Suppose $\{\underline{P}_i: i \geq 1\}$ is an increasing sequence of lower previsions on \mathcal{K} , so $\underline{P}_i \geq \underline{P}_j$ whenever $i \geq j$. If each \underline{P}_i avoids sure loss then the sequence converges pointwise to some \underline{P} on \mathcal{K} that avoids sure loss. If each \underline{P}_i is coherent then the limit \underline{P} is coherent.

Proof. Suppose each \underline{P}_i avoids sure loss. For every $X \in \mathcal{K}$, $\sup\{\underline{P}_i(X): i \geq 1\} \leq \sup X$ by property 2.4.7(a), so the increasing sequence of reals $\{\underline{P}_i(X): i \geq 1\}$ is bounded above and converges to some real $\underline{P}(X)$. The result then follows from Theorem 2.6.5. ♦

The corollary asserts that the supremum of an increasing sequence of coherent lower previsions is coherent. Compare with the lower envelope theorem, which asserts that the infimum of an arbitrary set of coherent lower previsions is coherent. The two results together imply that a lower limit of coherent lower previsions is coherent.

2.6.7 Corollary

Let $\{\underline{P}_i: i \geq 1\}$ be a sequence of coherent lower previsions on \mathcal{K} , and define \underline{P} on \mathcal{K} by $\underline{P}(X) = \liminf_{i \rightarrow \infty} \underline{P}_i(X)$. Then \underline{P} is a coherent lower prevision on \mathcal{K} .⁴

Proof. Define $\underline{Q}_n(X) = \inf\{\underline{P}_i(X): i \geq n\}$. Then $\underline{Q}_n(X)$ is finite since $\underline{P}_i(X) \geq \inf X$. So $\{\underline{Q}_n: n \geq 1\}$ is an increasing sequence of lower previsions that converges pointwise to \underline{P} . Each \underline{Q}_n is coherent by Theorem 2.6.3(b), so \underline{P} is coherent by Corollary 2.6.6. ♦

As a simple application of the last result, consider the following frequentist interpretation of coherent lower previsions.

2.6.8 Divergent relative frequencies

Consider a hypothetical infinite sequence of repetitions of an experiment with sample space Ω . Let ω_j denote the outcome of the j th repetition. For each $n \geq 1$, define the **sample-mean prevision** P_n for all gambles X by $P_n(X) = n^{-1} \sum_{j=1}^n X(\omega_j)$. Then $P_n(X+Y) = P_n(X) + P_n(Y)$ and $\sup X \geq P_n(X) \geq \inf X$, so P_n is a linear prevision on $\mathcal{L}(\Omega)$ by the results of section 2.3.6. When A is the indicator function of an event (defined in 2.7.1), $P_n(A)$ is just the relative frequency of the event in the first n outcomes. The ‘limiting relative frequencies’ in the infinite sequence of repetitions can then be described by the lower prevision \underline{P} , where $\underline{P}(X) = \liminf_{n \rightarrow \infty} P_n(X)$ for every gamble X . By Corollary 2.6.7, \underline{P} is a coherent lower prevision on $\mathcal{L}(\Omega)$. Its conjugate upper prevision is $\bar{P}(X) = \limsup_{n \rightarrow \infty} P_n(X)$. We have $\underline{P}(X) = \bar{P}(X)$ just when

$P_n(X)$ converges to a limit; this need not happen for any non-constant gambles X . This standard frequentist model is therefore a source of coherent, non-linear, lower previsions.⁵

2.7 Coherent probabilities

The usual approach in constructing a theory of probability is to start with axioms for probability and then define expectation, with the mathematical properties of linear prevision, as a derived concept. We regard prevision as more fundamental than probability. In this respect we follow de Finetti (1974), and there is additional motivation to do so in a theory of imprecise probabilities. That is because lower previsions may ‘contain more information’ than lower probabilities, so that it may not be possible to adequately express Your beliefs just in terms of upper and lower probabilities. That is illustrated by the example in section 2.7.3. Nevertheless, in some problems it will be both convenient and sufficient to assess upper and lower probabilities for events. We therefore regard upper and lower probability as an important special case of upper and lower prevision. A lower probability is just a lower prevision defined on a special type of domain that contains only indicator functions of events. The following notation simplifies the mathematical theory.

2.7.1 Notation

We use capitals A, B, C, D, A_j to denote subsets of Ω , called **events**. As is standard, \emptyset denotes the empty set, A^c denotes the set-theoretic complement of A , $A \cup B$ and $A \cap B$ respectively denote the union and intersection of A and B , $A \Delta B = (A \cap B^c) \cup (A^c \cap B)$ denotes their set-theoretic difference, and \supset or \subset denotes set inclusion. If A is an event, we use the same symbol A to denote its **indicator function** on domain Ω , defined by $A(\omega) = 1$ if $\omega \in A$ and $A(\omega) = 0$ if $\omega \in A^c$. An indicator function is just a 0–1 valued gamble. Thus we identify each event with a gamble, and regard the class of all events as a subset of $\mathcal{L}(\Omega)$.¹ This convention greatly simplifies the notation and should not lead to confusion: whether the symbol A denotes an event (as in $A \cap B = \emptyset$) or a gamble (as in $\underline{P}(A) = \frac{1}{2}$) should be clear from the context. Note, however, that the gamble $A - B$ corresponds to the event $A \cap B^c$ only when $A \supset B$, and the gamble $A + B$ corresponds to the event $A \cup B$ only when $A \cap B = \emptyset$. Generally, the events $\Omega, \emptyset, A^c, A \cap B, A \cup B$ and $A \Delta B$ can be identified with the gambles 1, 0, 1 – A , AB , $A + B - AB$ and $|A - B|$ respectively. Set inclusion, $A \supset B$, is identified with the ordering on gambles $A \geq B$.

2.7.2 Lower and upper probability

Let \mathcal{A} denote an arbitrary class of events, which we regard as a class of 0–1 valued gambles. When the lower prevision \underline{P} is defined on such a class \mathcal{A} , we call \underline{P} a **lower probability** on \mathcal{A} , and call $\underline{P}(A)$ the lower probability of event A . As before, $\underline{P}(A)$ is interpreted as a supremum price You are willing to pay for the gamble A , which pays 1 unit if event A occurs (and nothing otherwise). Thus $\underline{P}(A)$ is a supremum betting rate at which You are disposed to bet on A .

It is convenient to modify the earlier definition of the conjugate upper prevision \bar{P} , now called the **upper probability**, so that \bar{P} is defined on $\mathcal{A}^c = \{A^c : A \in \mathcal{A}\} = \{1 - A : A \in \mathcal{A}\}$ rather than on $-\mathcal{A} = \{-A : A \in \mathcal{A}\}$. (This enables us to restrict attention to events.²) Define \bar{P} on \mathcal{A}^c by $\bar{P}(A) = 1 - \underline{P}(A^c)$.

As before, $G(A) = A - \underline{P}(A)$ denotes the marginal gamble on A . Betting against A is equivalent to betting on A^c , and $G(A^c) = A^c - \underline{P}(A^c) = (1 - A) - (1 - \bar{P}(A)) = \bar{P}(A) - A$. Thus $\bar{P}(A)$ can be interpreted as Your infimum selling price for the gamble A , or as the infimum betting rate at which You will take bets on A , or as one minus the supremum betting rate at which You will bet against A . It is a simple consequence of avoiding sure loss that $\underline{P}(A) \leq \bar{P}(A)$ whenever both A and A^c are in \mathcal{A} .³ The imprecision of the lower probability \underline{P} is naturally measured by the values of $\bar{P}(A) - \underline{P}(A)$.⁴

We now give an example to show that beliefs cannot always be fully described through upper and lower probabilities.⁵

2.7.3 Lower probabilities may not determine lower previsions

Let $\Omega = \{a, b, c\}$ be a 3-point set. Define the probability mass functions $P_j = (P_j(a), P_j(b), P_j(c))$ for $1 \leq j \leq 4$, by $P_1 = (\frac{2}{3}, \frac{1}{3}, 0)$, $P_2 = (\frac{1}{3}, 0, \frac{2}{3})$, $P_3 = (\frac{2}{3}, 0, \frac{1}{3})$ and $P_4 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Each P_j has a unique extension to a linear prevision on $\mathcal{L}(\Omega)$, defined by $P_j(X) = \sum_{\omega \in \Omega} P_j(\omega)X(\omega)$. We will construct several different lower previsions as lower envelopes of these linear previsions, and show that the lower previsions yield the same lower probabilities when they are restricted to events.

First define \underline{P}_1 to be the lower envelope of the class $\{P_1, P_2\}$ of linear previsions, so $\underline{P}_1(X) = \min\{P_1(X), P_2(X)\}$ for all $X \in \mathcal{L}$. By Theorem 2.6.3, \underline{P}_1 is a coherent lower prevision. The restriction of \underline{P}_1 to events is therefore a coherent lower probability,⁶ and since Ω has only three elements this is determined by the upper and lower probabilities of singletons:

	$\{a\}$	$\{b\}$	$\{c\}$
\bar{P}_1	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
\underline{P}_1	$\frac{1}{3}$	0	0

2.7 COHERENT PROBABILITIES

Now consider P_3 . Note that $\underline{P}_1(\{\omega\}) \leq P_3(\omega) \leq \bar{P}_1(\{\omega\})$ for each $\omega \in \Omega$. Hence P_3 dominates \underline{P}_1 on the class of all events. Defining \underline{P}_2 to be the lower envelope of $\{P_1, P_2, P_3\}$, \underline{P}_2 therefore agrees with \underline{P}_1 on events. (Including P_3 does not reduce any of the minima which define lower probabilities of events.) Similarly, define \underline{P}_3 and \underline{P}_4 as lower envelopes of $\{P_1, P_2, P_4\}$ and $\{P_1, P_2, P_3, P_4\}$ respectively. Note that P_4 also dominates \underline{P}_1 on events. It follows that $\underline{P}_1, \underline{P}_2, \underline{P}_3, \underline{P}_4$ are coherent lower previsions that all agree on events.

To see that the four lower previsions are different, it suffices to consider the gamble $X = \{a\} - \{b\}$. Then $P_1(X) = P_2(X) = \frac{1}{3}$, $P_3(X) = \frac{2}{3}$, $P_4(X) = 0$. Hence $\underline{P}_1(X) = \bar{P}_1(X) = \frac{1}{3}$, $\underline{P}_2(X) = \frac{1}{3}$, $\bar{P}_2(X) = \frac{2}{3}$, $P_3(X) = 0$, $\bar{P}_3(X) = \frac{1}{3}$, $P_4(X) = 0$, $\bar{P}_4(X) = \frac{2}{3}$. Previsions for X are precise under \underline{P}_1 but highly imprecise under \underline{P}_4 . The four lower previsions can give rise to quite different behaviour. For example, the gamble $X - \frac{1}{3}$ is equivalent to the zero gamble under \underline{P}_1 , marginally desirable under \underline{P}_2 , marginally undesirable under \underline{P}_3 , and its desirability is indeterminate under \underline{P}_4 .

The probability mass functions P_j in this example can be represented as points in the probability simplex (equilateral triangle), where $P_j(\omega)$ is the perpendicular distance of the point from the side opposite vertex ω . (See section 4.2 for discussion of this graphical representation.)

The additive probabilities that dominate the lower probability \underline{P}_1 on events are just those in the convex hull of $\{P_1, P_2, P_3, P_4\}$, represented by a parallelogram in Figure 2.7.3. The upper and lower probabilities of events determine lines parallel to the sides of the equilateral triangle. It is easy to see that if P_1 and P_2 are both included in the region bounded by such lines, then so must be P_3 and P_4 . This means that You cannot distinguish between the lower previsions $\underline{P}_1, \underline{P}_2, \underline{P}_3, \underline{P}_4$ merely by specifying the upper and

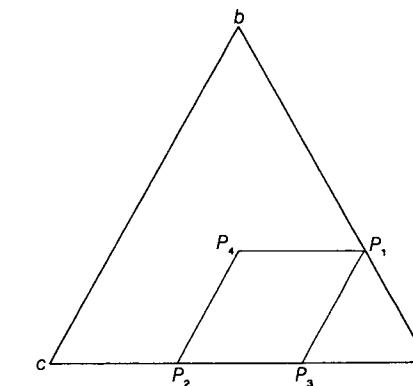


Figure 2.7.3 Simplex representation for Example 2.7.3

lower probabilities of events.⁷ But You can do so by specifying upper and lower previsions for other gambles, which bound the class of dominating additive probabilities by other straight lines. When $X = \{a\} - \{b\}$, for example, the assessments $\underline{P}(X) = \bar{P}(X) = \frac{1}{3}$ restrict the dominating additive probabilities to lie on the line joining P_1 and P_2 , which determines the lower prevision \underline{P}_1 .

Next consider coherence of lower probabilities. By Lemma 2.5.4(b), a lower probability \underline{P} on \mathcal{A} is coherent if and only if, whenever m and n are non-negative integers and $A_0, A_1, \dots, A_n \in \mathcal{A}$, there is $\omega \in \Omega$ such that $\sum_{j=1}^n G(A_j)(\omega) \geq mG(A_0)(\omega)$. Provided Ω has finite cardinality n , coherence can be verified, using the theorem in Appendix A3, by solving sets of n linear equations for n or fewer unknowns.⁸

The basic properties of coherent upper and lower probabilities are listed next, for convenient reference. Except for the slight modification to the definition of the conjugate upper probability, these properties are special cases of the properties listed in section 2.6.1. It is a useful exercise to derive them directly from the coherence condition above.

2.7.4 Properties of coherent upper and lower probabilities

Suppose \mathcal{A} is a class of events, \underline{P} is a coherent lower probability on domain \mathcal{A} and \bar{P} is the conjugate upper probability, $\bar{P}(A) = 1 - \underline{P}(A^\complement)$ whenever $A^\complement \in \mathcal{A}$. The following properties hold whenever their terms are defined:

- (a) $0 \leq \underline{P}(A) \leq \bar{P}(A) \leq 1$
- (b) $\underline{P}(\emptyset) = \bar{P}(\emptyset) = 0$, $\underline{P}(\Omega) = \bar{P}(\Omega) = 1$
- (c) if $A \supset B$ then $\underline{P}(A) \geq \underline{P}(B)$ and $\bar{P}(A) \geq \bar{P}(B)$
- (d) $\underline{P}(A \cup B) \leq \underline{P}(A) + \bar{P}(B)$, $\bar{P}(A \cup B) \leq \bar{P}(A) + \bar{P}(B)$
- (e) if $A \cap B = \emptyset$ then $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$ and $\bar{P}(A \cup B) \geq \bar{P}(A) + \bar{P}(B)$
- (f) $\underline{P}(A) + \underline{P}(B) \leq 1 + \underline{P}(A \cap B)$, if $A \cup B = \Omega$ then $\bar{P}(A) + \bar{P}(B) \geq 1 + \bar{P}(A \cap B)$
- (g) $|\underline{P}(A) - \underline{P}(B)| \leq \bar{P}(A \Delta B)$, $|\bar{P}(A) - \bar{P}(B)| \leq \bar{P}(A \Delta B)$
- (h) $\underline{P}(A) + \underline{P}(B) \leq \underline{P}(A \cup B) + \bar{P}(A \cap B) \leq \bar{P}(A) + \bar{P}(B)$
 $\underline{P}(A) + \underline{P}(B) \leq \bar{P}(A \cup B) + \underline{P}(A \cap B) \leq \bar{P}(A) + \bar{P}(B)$
- (i) if $\bar{P}(A_n \Delta A) \rightarrow 0$ as $n \rightarrow \infty$ then $\underline{P}(A_n) \rightarrow \underline{P}(A)$ and $\bar{P}(A_n) \rightarrow \bar{P}(A)$

$$(j)^9 \quad \bar{P}\left(\bigcup_{j=1}^n A_j\right) \leq \sum_{j=1}^n \bar{P}(A_j), \text{ if } A_1, A_2, \dots \text{ are disjoint then}$$

$$\underline{P}\left(\bigcup_{j=1}^n A_j\right) \geq \sum_{j=1}^n \underline{P}(A_j) \text{ and } \underline{P}\left(\bigcup_{j=1}^{\infty} A_j\right) \geq \sum_{j=1}^{\infty} \underline{P}(A_j)$$

- (k) $(\forall \lambda_1, \dots, \lambda_n \geq 0 \text{ and } \mu \in \mathbb{R})$

$$\text{if } B \geq \sum_{j=1}^n \lambda_j A_j + \mu \text{ then } \underline{P}(B) \geq \sum_{j=1}^n \lambda_j \underline{P}(A_j) + \mu,$$

$$\text{if } B \leq \sum_{j=1}^n \lambda_j A_j + \mu \text{ then } \bar{P}(B) \leq \sum_{j=1}^n \lambda_j \bar{P}(A_j) + \mu$$

$$(l) \text{ if } \lambda_1, \dots, \lambda_m, \beta_1, \dots, \beta_m \geq 0 \text{ and } \sum_{j=1}^n \lambda_j A_j \leq \sum_{i=1}^m \beta_i B_i \text{ then } \sum_{j=1}^n \lambda_j \underline{P}(A_j) \leq \beta_1 \underline{P}(B_1) + \sum_{i=2}^m \beta_i \bar{P}(B_i), \text{ and } \lambda_1 \bar{P}(A_1) + \sum_{j=2}^n \lambda_j \underline{P}(A_j) \leq \sum_{i=1}^m \beta_i \bar{P}(B_i).$$

It does not seem possible to characterize coherence of lower probabilities through simple axioms such as (a)–(h) above, even when the domain \mathcal{A} is the class of all events. (Compare with the simple axioms P1–P3 which characterize coherence on linear spaces.) Often the easiest way to show that a given lower probability is coherent is to show that it is the restriction to events of some lower revision defined on a linear space, whose coherence can be directly verified.

The lower probabilities \underline{P} in the next two examples satisfy the basic properties (a)–(h), but are not coherent. In the first example \underline{P} does not even avoid sure loss. The second example shows that the basic properties together with avoiding sure loss are still not sufficient for coherence.

2.7.5 Lower probabilities that incur sure loss

Let Ω be a 7-point space, $\Omega = \{0, 1, 2, 3, 4, 5, 6\}$. Let \oplus denote addition modulo 7. Define the class of events $\mathcal{I} = \{B_j : 0 \leq j \leq 6\}$ where $B_j = \{j, j \oplus 2, j \oplus 3\}$. Any two distinct sets in \mathcal{I} have exactly one point in common, and each point of Ω belongs to exactly three sets in \mathcal{I} . Define \underline{P} for every event A by $\underline{P}(\Omega) = 1$, $\underline{P}(A) = \frac{1}{2}$ if $A \neq \Omega$ but A contains some B_j , and $\underline{P}(A) = 0$ otherwise. Let \bar{P} be the conjugate upper probability. It can be verified that \underline{P} and \bar{P} satisfy all the properties (a) to (h) of section 2.7.4. For example, the sub-additivity property $\bar{P}(A \cup B) \leq \bar{P}(A) + \bar{P}(B)$ holds because $\bar{P}(A)$ takes only the values $\frac{1}{2}$ and 1 when A is non-empty. The super-additivity property, $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$ when $A \cap B = \emptyset$, holds because no two members of \mathcal{I} are disjoint. But \underline{P} incurs sure loss, since $\sum_{j=0}^6 B_j = 3$ but $\underline{P}(B_j) = \frac{1}{2}$ for $0 \leq j \leq 6$, so that $\sum_{j=0}^6 G(B_j) = 3 - \frac{7}{2} = -\frac{1}{2}$.

2.7.6 Incoherent lower probabilities that avoid sure loss

Let $\Omega = \{a, b, c, d\}$. Define \underline{P} for all events A by $\underline{P}(\Omega) = 1$, $\underline{P}(A) = \frac{1}{2}$ if A contains a and one or two other points, and $\underline{P}(A) = 0$ otherwise. Then \underline{P} is

dominated by the uniform probability distribution on events, so \underline{P} avoids sure loss by Theorem 2.6.3(a). Because \underline{P} essentially takes only the values 0 and $\frac{1}{2}$, properties (a) to (h) of section 2.7.4 are again easy to verify, as in the previous example. But \underline{P} is not coherent.¹⁰ To see that, let $A = \{a\}$, $B = \{a, b\}$, $C = \{a, c\}$, $D = \{a, d\}$, so that $\underline{P}(A) = 0$, $\underline{P}(B) = \underline{P}(C) = \underline{P}(D) = \frac{1}{2}$. Since $B + C + D - 2A = 1$, we have $G(B) + G(C) + G(D) - 2G(A) = 1 - \frac{3}{2} = -\frac{1}{2}$. This violates coherence. To achieve coherence, $\underline{P}(A)$ must be increased to $\frac{1}{4}$, since the even-money bets on each of B , C , D effectively require You to pay up to $\frac{1}{4}$ for A .

2.8 Linear previsions and additive probabilities

Our definition of coherence allows Your infimum selling price for a gamble X , $\bar{P}(X)$, to be strictly larger than Your supremum buying price for X , $\underline{P}(X)$. That distinguishes the present theory from that of de Finetti (1974), which is also based on a notion of coherence but requires that $\underline{P}(X) = \bar{P}(X)$ for all gambles X . When the lower and upper previsions coincide and are coherent we call them **linear previsions**, and denote the common value of $\underline{P}(X)$ and $\bar{P}(X)$ by $P(X)$. The term ‘linear’ is appropriate because any linear prevision agrees on its domain with some linear functional defined on $\mathcal{L}(\Omega)$. The latter can in turn be identified with a finitely additive probability defined on all subsets of Ω .¹¹

Linear previsions play an important role in our theory, as both a mathematical model for determinate uncertainty and a means of constructing coherent lower previsions: every coherent lower prevision is a lower envelope of some class of linear previsions. The following definition of linear previsions reduces to agreement of upper and lower previsions when $\mathcal{K} = -\mathcal{K}$ but applies more generally, to an arbitrary domain \mathcal{K} .

2.8.1 Definition of linear prevision

Suppose that P is a real-valued function defined on a class of gambles \mathcal{K} . Let $G(X)$ denote the marginal gamble $X - P(X)$. Then P is called a **linear prevision** on \mathcal{K} if $\sup[\sum_{j=1}^n G(X_j) - \sum_{j=1}^m G(Y_j)] \geq 0$ whenever m and n are non-negative integers and $X_1, \dots, X_n, Y_1, \dots, Y_m$ are in \mathcal{K} .

Compare this definition with the general definitions of avoiding sure loss (2.4.1) and coherence (2.5.1). The prevision $P(X)$ is interpreted as Your ‘fair price’ for X , meaning that You are both willing to buy X for any price less than $P(X)$ and to sell X for any price greater than $P(X)$. When $X \in \mathcal{K}$, the gambles $G(X) = X - P(X)$ and $-G(X) = P(X) - X$ are each marginally desirable. Definition 2.8.1 then says that Your fair prices must ‘avoid sure loss’, in the previous sense that there is no finite sum of desirable gambles

that is uniformly negative. Provided You do have fair prices, it is a requirement of rationality that these should be linear previsions. But it is not obvious that You should have fair prices for every gamble. The arguments for this version of the bayesian dogma of precision will be examined in section 5.7. We merely note here that it is a strong assumption which needs careful justification.

Every linear prevision is a coherent lower prevision, by taking each $Y_j = X_0$ in Definition 2.8.1, and also a coherent upper prevision, by taking each $X_j = Y_0$. As in Lemma 2.4.4(a), the definition of linear prevision is equivalent to the condition: $\sup \sum_{j=1}^n \lambda_j G(X_j) \geq 0$ whenever $n \geq 1$, $X_j \in \mathcal{K}$ and $\lambda_j \in \mathbb{R}$ for $1 \leq j \leq n$.² (Here the λ_j are any real numbers, whereas in Lemma 2.4.4 they were restricted to be non-negative.)

The next results show that the definition of linear prevision simplifies in the expected ways when the domain \mathcal{K} satisfies $\mathcal{K} = -\mathcal{K}$, or is closed under addition of gambles, or is a linear space. The first result shows that the general definition of linear prevision agrees with the definition given (for linear spaces) in section 2.3. A lower prevision is linear if and only if it is coherent and self-conjugate. **Self-conjugacy** means that when P is regarded as a lower prevision, it agrees with its conjugate upper prevision. So Your supremum buying price is equal to Your infimum selling price for every gamble.

2.8.2 Theorem

Let P be a lower prevision on domain \mathcal{K} where $\mathcal{K} = -\mathcal{K}$. Then P is a linear prevision on \mathcal{K} if and only if it avoids sure loss and satisfies the self-conjugacy property $P(X) = -P(-X)$ for all $X \in \mathcal{K}$. Linearity is equivalent also to coherence plus self-conjugacy.

Proof. Suppose P is a linear prevision on \mathcal{K} . By setting each $Y_j = X_0$ in Definition 2.8.1, P is a coherent lower prevision, and so P also avoids sure loss. By property 2.4.7(a), $P(X) \leq -P(-X)$ for all $X \in \mathcal{K}$. Take $n = 0$, $m = 2$, $Y_1 = X$, $Y_2 = -X$ in 2.8.1 to show $P(X) \geq -P(-X)$.

Conversely, suppose P avoids sure loss and $(\forall X \in \mathcal{K})P(X) = -P(-X)$. Given $X_1, \dots, X_n, Y_1, \dots, Y_m \in \mathcal{K}$, write $X_{n+j} = -Y_j$ for $1 \leq j \leq m$, so $G(X_{n+j}) = X_{n+j} - P(X_{n+j}) = -Y_j - P(-Y_j) = -Y_j + P(Y_j) = -G(Y_j)$. Then $\sup[\sum_{j=1}^n G(X_j) - \sum_{j=1}^m G(Y_j)] = \sup[\sum_{j=1}^{n+m} G(X_j)] \geq 0$ since P avoids sure loss. Thus P is a linear prevision on \mathcal{K} . ◆

Say that P avoids sure gain if $\inf \sum_{j=1}^n G(X_j) \leq 0$ whenever $X_1, \dots, X_n \in \mathcal{K}$. In that case there is no finite sum of marginal gambles that is uniformly positive. By taking $n = 0$ in Definition 2.8.1, every linear prevision avoids sure gain. (Consequently the betting rates set by bookmakers, which are

chosen in order to achieve a sure gain, cannot be linear previsions or additive probabilities.) It can be seen from the above theorem that, when $\mathcal{K} = -\mathcal{K}$, a lower prevision is linear if and only if it simultaneously avoids sure loss and avoids sure gain.³

The next result should be compared with the result at the end of section 2.5 that, under the same conditions on \mathcal{K} , coherence is equivalent to axioms P2', P3 and P4. Its proof is left as an exercise.

2.8.3 Theorem

Suppose P is defined on a domain \mathcal{K} which is closed under addition of gambles and contains the zero gamble. Then P is a linear prevision on \mathcal{K} if and only if it satisfies the two axioms, for all X, Y in \mathcal{K} ,

$$(P0) \quad P(X + Y) = P(X) + P(Y) \quad (\text{linearity})$$

$$(P4) \quad \text{if } \mu \in \mathbb{R} \text{ and } X \geq Y + \mu \text{ then } P(X) \geq P(Y) + \mu \quad (\text{monotonicity}).$$

The next result characterizes the linear previsions on linear spaces of gambles. It should be compared with the result (Theorem 2.5.5) that coherence of lower previsions is equivalent to the axioms P1, P2, P3 introduced in section 2.3.

2.8.4 Theorem

Suppose P is defined on a linear space of gambles \mathcal{K} . Then P is a linear prevision if and only if it satisfies the linearity axiom P0 and the ‘accepting sure gain’ axiom:

$$(P1) \quad P(X) \geq \inf X \quad \text{when } X \in \mathcal{K}.$$

*Proof.*⁴ A linear prevision satisfies P0, by Definition 2.8.1 with $X_1 = X$, $X_2 = Y$, $Y_1 = X + Y$ and $X_1 = X + Y$, $Y_1 = X$, $Y_2 = Y$, and satisfies P1 by taking $Y_1 = X$ in 2.8.1. Conversely, suppose P0 and P1 hold. Given $X_1, \dots, X_n, Y_1, \dots, Y_m \in \mathcal{K}$, let $Z = \sum_{j=1}^m Y_j - \sum_{j=1}^n X_j$, so $Z \in \mathcal{K}$. By P0, $P(Z) = \sum_{j=1}^m P(Y_j) - \sum_{j=1}^n P(X_j)$. By P1, $P(Z) \geq \inf Z$. Hence

$$\sup \left[\sum_{j=1}^n X_j - \sum_{j=1}^m Y_j \right] = \sup(-Z) = -\inf Z \geq -P(Z) = \sum_{j=1}^n P(X_j) - \sum_{j=1}^m P(Y_j),$$

which establishes 2.8.1. ◆

2.8.5 Corollary

Suppose P is defined on a linear space of gambles, \mathcal{K} , which contains constant gambles. Then P is a linear prevision on \mathcal{K} if and only if it satisfies

the four axioms P0 (linearity), P2 (positive homogeneity),⁵

(P5) if $X \in \mathcal{K}$ and $X \geq 0$ then $P(X) \geq 0$ (positivity)

(P6) $P(1) = 1$ (unit norm).

(These axioms define the positive linear functionals with unit norm on \mathcal{K}).⁶

Proof. Suppose P is a linear prevision. P0 and P1 hold by Theorem 2.8.4, and P1 implies P5. Also P is coherent by Theorem 2.8.2, so P2 holds by Theorem 2.5.5 and P6 holds by property 2.6.1(b). Conversely, suppose P satisfies the four axioms. First show that $P(\mu) = \mu$ when $\mu \in \mathbb{R}$. This is true if $\mu > 0$, by P2 and P6. It extends to all $\mu \in \mathbb{R}$ using $P(-\mu) + P(\mu) = P(0) = 0$, by P0. Given $X \in \mathcal{K}$, write $\mu = \inf X$ and use P5 to show $P(X - \mu) \geq 0$ so that $P(X) = P(X - \mu) + P(\mu) \geq \mu$ by P0. This establishes P1, so P is a linear prevision by 2.8.4. ◆

In many problems, the simplest way to establish coherence of a given lower prevision is to show that it is a lower envelope of linear previsions, and invoke the following corollary of the lower envelope theorem (2.6.3).

2.8.6 Corollary

(a) If \underline{P} is a lower prevision with domain \mathcal{K} that is dominated by a linear prevision on \mathcal{K} , then \underline{P} avoids sure loss.

(b) If \underline{P} is the lower envelope of a non-empty class of linear previsions defined on \mathcal{K} , then \underline{P} is a coherent lower prevision on \mathcal{K} .

In Chapter 3 we will prove that the converse holds: \underline{P} avoids sure loss if and only if it is dominated by a linear prevision, and \underline{P} is coherent if and only if it is the lower envelope of a class of linear previsions.

2.8.7 Additive probabilities

Next we introduce additive probability as a special case of linear prevision. A real-valued function P defined on a class of events \mathcal{A} is called an **additive probability** on \mathcal{A} when P is a linear prevision on the corresponding class of indicator functions (as in section 2.7.1). We show below that this definition is equivalent to the usual definition of finitely additive probability when \mathcal{A} is a field of events. (Throughout the book, ‘additive’ will mean ‘finitely additive’ and carries no implication of countable additivity.) The additive probability $P(A)$ can be interpreted as a fair betting rate: You are willing to bet on A at rates arbitrarily close to $P(A)$, and also willing to bet against A at rates arbitrarily close to $1 - P(A)$.

Using Definition 2.8.1 and Lemma 2.4.4(b), P is an additive probability on \mathcal{A} if and only if, whenever m and n are non-negative integers and

$A_1, \dots, A_n, B_1, \dots, B_m \in \mathcal{A}$, there is $\omega \in \Omega$ such that

$$\sum_{j=1}^n A_j(\omega) - \sum_{j=1}^m B_j(\omega) \geq \sum_{j=1}^n P(A_j) - \sum_{j=1}^m P(B_j).$$

The following results show that this condition simplifies when \mathcal{A} is closed under complementation, or when \mathcal{A} is a field of events.

2.8.8 Lemma

Suppose P is defined on a class of events \mathcal{A} that is closed under complementation. Then P is an additive probability on \mathcal{A} if and only if P avoids sure loss (as a lower probability) and $P(A) + P(A^c) = 1$ for all $A \in \mathcal{A}$.

Proof. If P is an additive probability then it avoids sure loss. Hence $P(A) + P(A^c) \leq 1$. To show $P(A) + P(A^c) \geq 1$, take $n = 0$, $m = 2$, $B_1 = A$, $B_2 = A^c$ in the above condition.

Conversely, given $A_1, \dots, A_n, B_1, \dots, B_m \in \mathcal{A}$, write $A_{n+j} = B_j^c$ for $1 \leq j \leq m$, so that $P(A_{n+j}) = 1 - P(B_j)$. Since P avoids sure loss, there is $\omega \in \Omega$ such that $\sum_{j=1}^n A_j(\omega) - \sum_{j=1}^m B_j(\omega) = \sum_{j=1}^{n+m} A_j(\omega) - m \geq \sum_{j=1}^{n+m} P(A_j) - m = \sum_{j=1}^n P(A_j) - \sum_{j=1}^m P(B_j)$. Thus P is an additive probability. \diamond

Here $P(A) + P(A^c) = 1$ is a self-conjugacy property; when an additive probability P is regarded as a lower probability, the conjugate upper probability agrees with P . Thus a lower probability that is self-conjugate and avoids sure loss must be additive.⁷

A class of events \mathcal{A} is a **field** when (i) $\Omega \in \mathcal{A}$, (ii) $A^c \in \mathcal{A}$ whenever $A \in \mathcal{A}$, and (iii) $A \cup B \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$. Every field is closed under finite unions and intersections and contains the empty set. Additivity on a field is equivalent to the familiar axioms for finitely additive probability.

2.8.9 Theorem

Suppose P is defined on a field of events \mathcal{A} . Then P is an additive probability on \mathcal{A} if and only if it satisfies the three axioms, for all $A, B \in \mathcal{A}$,

- (a) $P(A) \geq 0$ (non-negativity)
- (b) $P(\Omega) = 1$ (normalization)
- (c) $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$ (finite additivity).

Proof. If P is an additive probability then it is a coherent lower probability and is self-conjugate, so (a), (b), (c) follow from properties (a), (b), (d), (e) of section 2.7.4. For the converse, suppose the three axioms hold. Then $P(A) + P(A^c) = 1$, by taking $B = A^c$ in (c) and using (b). By Lemma 2.8.8 it suffices to prove that P avoids sure loss. Given $A_1, \dots, A_n \in \mathcal{A}$, let B_1, \dots, B_m

be the atoms of the field generated by $\{A_1, \dots, A_n\}$, i.e. the distinct non-empty sets of the form $\bigcap_{j=1}^n C_j$ where each C_j is either A_j or A_j^c . Each B_k is in \mathcal{A} since \mathcal{A} is a field, $\{B_1, \dots, B_m\}$ is a partition of Ω , and $\sum_{k=1}^m P(B_k) = 1$ by (b) and (c). Each A_j can be written as a finite union of disjoint sets B_k , and $P(A_j) = \sum_{B_k \subseteq A_j} P(B_k)$ by (c). Let n_k denote the number of sets A_1, \dots, A_n which contain B_k . Note that $\sum_{j=1}^n A_j(\omega) = n_k$ when $\omega \in B_k$. Hence $\sup \sum_{j=1}^n A_j = \max \{n_k : 1 \leq k \leq m\} = \sum_{k=1}^m P(B_k) \max \{n_k : 1 \leq k \leq m\} \geq \sum_{k=1}^m P(B_k) n_k = \sum_{k=1}^m \sum_{A_j \supseteq B_k} P(B_k) = \sum_{j=1}^n \sum_{B_k \subseteq A_j} P(B_k) = \sum_{j=1}^n P(A_j)$. Thus $\sup \sum_{j=1}^n (A_j - P(A_j)) \geq 0$, and P avoids sure loss. \diamond

2.8.10 Extension of additive probability to linear prevision

It is well known, and will be proved in section 3.2, that any additive probability defined on a field of events \mathcal{A} has a unique extension to a positive linear functional with unit norm (linear prevision) defined on the class $\mathcal{K}(\mathcal{A})$ of \mathcal{A} -measurable gambles. Hence there is a one-to-one correspondence between additive probabilities on \mathcal{A} and linear previsions on $\mathcal{K}(\mathcal{A})$.

We will often make use of Corollary 2.8.6 by first defining a class of additive probabilities on some domain \mathcal{A} , uniquely extending these to linear previsions on $\mathcal{K}(\mathcal{A})$, and finally defining a coherent lower prevision on $\mathcal{K}(\mathcal{A})$ to be the lower envelope of the class of linear previsions. An example of this construction was given in section 2.7.3.

Because of the one-to-one correspondence between linear previsions and additive probabilities, Bayesian theories can take probability as the fundamental concept, and define prevision or expectation as an integral with respect to a probability measure. Since an additive probability defined on all subsets of Ω uniquely determines a linear prevision on all gambles in $\mathcal{L}(\Omega)$, You can fully specify Your beliefs about Ω through (precise) probabilities of events, without needing to directly assess previsions for more complicated gambles.⁸ Compare this with the case of non-additive lower probabilities illustrated in Example 2.7.3. Beliefs concerning Ω cannot always be fully specified through lower probabilities of events, as a coherent lower probability defined on all events may have many different extensions to a coherent lower prevision on $\mathcal{L}(\Omega)$.

2.8.11 Axioms of precision

In this section we have introduced several ‘axioms of precision’ to characterize linear previsions and additive probabilities. These include the two self-conjugacy axioms, $P(X) = -P(-X)$ and $P(A) + P(A^c) = 1$, the linearity axiom P0, and the finite additivity axiom.⁹ In effect, these axioms require

that previsions be precise, in the sense that lower previsions agree with their conjugate upper previsions. In the presence of these axioms, coherence is equivalent to avoiding sure loss.

Unlike the coherence axioms P1–P3, however, the axioms of precision are not compelling rationality criteria. If You have little information that can be used in assessing the probability of an event A , then You cannot be expected to assess a precise probability for A . For example, let A be the event that an earthquake with intensity 7 or greater on the Modified Mercalli scale is experienced in New Zealand in 1999. In Your current state of ignorance about the frequency of earthquakes in New Zealand and about the Modified Mercalli scale, are You prepared to assess a fair betting rate $P(A)$ and accept bets on or against A at rates arbitrarily close to $P(A)$, as required by the axioms of precision?¹⁰

2.9 Examples of coherent previsions

In this section we introduce some of the examples of coherent upper and lower previsions that will be elaborated in later chapters. The aims here are to define the upper and lower previsions, to prove they are coherent, and to indicate the kinds of applications in which they are useful models. Other examples, not discussed in this section, include (a) examples on a 3-point space Ω , for which upper and lower previsions have a simple graphical representation, used throughout Chapter 4 (see also Appendix I and Example 2.7.3); (b) the horse-racing example in Appendix C; (c) various statistical models discussed in Chapters 5, 7, 8 and 9.

2.9.1 Vacuous previsions

The vacuous lower and upper previsions were defined in section 2.3.7, by $\underline{P}(X) = \inf X$ and $\bar{P}(X) = \sup X$ for all gambles $X \in \mathcal{L}(\Omega)$. Their restrictions to events are called the **vacuous upper and lower probabilities**, $\underline{P}(\emptyset) = \bar{P}(\emptyset) = 0$, $\underline{P}(\Omega) = \bar{P}(\Omega) = 1$, $\underline{P}(A) = 0$ and $\bar{P}(A) = 1$ for all non-trivial events A .

It was noted in section 2.3.7 that the vacuous previsions are coherent, maximize the degree of imprecision $\bar{P}(X) - \underline{P}(X)$ for every gamble X amongst all coherent previsions, and make minimal claims about behavioural dispositions. They seem to be the only reasonable models for ‘complete ignorance’ concerning the possibility space Ω . Note that they do have the invariance properties one would expect of a model for complete ignorance. In particular, they are invariant under refinements or coarsenings of the space Ω ; the probability of an event does not depend on the space Ω in which it is represented. For example, if You are interested in the event that England win a specific football match against Scotland, the vacuous upper

and lower probabilities are one and zero whether You list the possible outcomes as $\Omega_1 = \{\text{win, not win}\}$, $\Omega_2 = \{\text{win, draw, lose}\}$, or in more detail (by specifying the score, for instance).

Contrast this with the behaviour of uniform probability distributions, which are used by many Bayesians to model complete ignorance. A uniform distribution has strong implications for behaviour (buying prices agree with selling prices for all gambles), and the probability of an event depends on the space Ω in which it is embedded. (In the example, the probability of England winning is $\frac{1}{2}$ under the uniform distribution on Ω_1 , $\frac{1}{3}$ under the uniform distribution on Ω_2).¹¹

The vacuous previsions are really rather trivial models. That seems appropriate for models of ‘complete ignorance’, which is a rather trivial state of uncertainty. On the other hand, one cannot expect such models to be very useful in practical problems, notwithstanding their theoretical importance. If the vacuous previsions are used to model prior beliefs about a statistical parameter, for instance, they give rise to vacuous posterior previsions. (See section 7.4.1 for discussion.) However, prior previsions that are close to vacuous, and make nearly minimal claims about prior beliefs, can lead to reasonable posterior previsions. (See section 5.3 for simple examples.)

2.9.2 Linear–vacuous mixtures

Vacuous previsions are coherent and so are linear previsions, so we can construct new coherent previsions by forming convex combinations of the two. If P_0 is a linear prevision on \mathcal{K} and $0 < \delta < 1$, define the **linear–vacuous mixture** (P_0, δ) on \mathcal{K} to be the convex combination of P_0 with the vacuous prevision, defined by $\underline{P}(X) = (1 - \delta)P_0(X) + \delta \inf X$. This is a coherent lower prevision by Theorem 2.6.4. The conjugate upper prevision is $\bar{P}(X) = (1 - \delta)P_0(X) + \delta \sup X$, so that the imprecision $\bar{P}(X) - \underline{P}(X) = \delta(\sup X - \inf X)$ is a fraction δ of the imprecision under the vacuous model.²

Linear–vacuous mixtures can be used by cautious Bayesians who assess a linear prevision P_0 but have incomplete confidence in their assessment. Specifically, consider the event D that You made a mistake in the analysis which led to P_0 . Suppose that You assess the precise probability of D to be δ , and Your beliefs about Ω conditional on D are vacuous. Then it is necessary for coherence that Your lower previsions are

$$\underline{P}(X) = P(D)\underline{P}(X|D) + P(D^c)\underline{P}(X|D^c) = \delta \inf X + (1 - \delta)P_0(X),$$

which is just the linear–vacuous mixture.

An alternative interpretation is that, after You buy a gamble X for price $\underline{P}(X)$, You must pay an additional tax of $\delta(X(\omega) - \inf X)$, which is

a fraction δ of the amount by which Your reward exceeds the minimum possible reward.³ Your gain from the transaction, after tax, is then $X - \underline{P}(X) - \delta(X - \inf X) = (1 - \delta)X + \delta \inf X - \underline{P}(X)$. If $P_0(X)$ is Your fair price for X , the transaction will be marginally desirable when the buying price is $\underline{P}(X) = (1 - \delta)P_0(X) + \delta \inf X$, the linear-vacuous mixture.

The upper and lower probabilities defined by the model have the simple forms $\underline{P}(A) = (1 - \delta)P_0(A)$ and $\bar{P}(A) = (1 - \delta)P_0(A) + \delta$ for all non-trivial events A , so that the degree of imprecision $\bar{P}(A) - \underline{P}(A) = \delta$ is constant over events. Thus δ can be interpreted as a measure of imprecision, or as a buying price deflation rate for betting on events. (Your fair rate $P_0(A)$ for betting on A is reduced by a proportion δ to get Your actual betting rate $\underline{P}(A)$.)

The linear-vacuous mixture (P_0, δ) is the lower envelope of the class of all linear previsions $(1 - \delta)P_0 + \delta P$, where P is any linear prevision. This class may be regarded as a ‘neighbourhood’ of P_0 . It has been used in Bayesian and frequentist studies of robustness.⁴

2.9.3 Pari-mutuel models, bookmaking and life insurance

A different kind of neighbourhood of a linear prevision can be defined by considering the pari-mutuel betting system used at racetracks. Let A_i be the event that horse i wins the race of interest, let $P_0(A_i)$ be the proportion of total stakes bet on horse i , and let τ be the proportion of total stakes deducted by the system operators ($0 < \tau < 1$). A unit bet on horse i will return $(1 - \tau)/P_0(A_i)$ if horse i wins.⁵ In effect, the pari-mutuel selling price for the gamble A_i (which pays one unit if i wins) is $\bar{P}(A_i) = P_0(A_i)/(1 - \tau) = (1 + \delta)P_0(A_i)$, where we define $\delta = \tau/(1 - \tau)$. Here δ is a selling price inflation rate; compare with the linear-vacuous models, for which δ is a buying price deflation rate.

This upper probability \bar{P} avoids sure loss because it dominates the additive probability P_0 . Consequently there is no way for a punter to construct a system of ‘win’ bets which guarantees him a profit.⁶ However, such systems are possible if other sorts of bets are allowed; see Appendix C for the case of ‘win’ and ‘place’ bets.

The upper probability \bar{P} is incoherent if $P_0(A_i) > 1 - \tau$ for some i , so that $\bar{P}(A_i) > 1$. In that case a bet on horse i is sure to produce a loss, and punters will tend to force the price $\bar{P}(A_i)$ downwards by betting on other horses. Coherence is therefore rarely violated by the final racetrack odds. To ensure coherence we will redefine \bar{P} to be $\bar{P}(A) = \min\{(1 + \delta)P_0(A), 1\}$. To see that this upper probability is coherent, when defined for any subsets A of finite Ω , note that it is the upper envelope of the class of additive probabilities of the form $P(\{\omega_j\}) = (1 + \delta)P_0(\{\omega_j\})$ for $j < m$, $P(\{\omega_j\}) = 0$ for $j > m$, for all orderings $\omega_1, \dots, \omega_n$ of Ω .⁷

The same model \bar{P} may be appropriate for the betting rates that are set by bookmakers, which also have the properties of coherent upper probabilities. In this case $P_0(A)$ is the fraction of stakes that the bookmaker anticipates will be bet on event A and τ is the fraction of total stakes set aside to cover profits, errors in anticipating the betting market, and any additional taxes and expenses. Provided the bookmaker’s anticipations are approximately correct, so that the actual fractions of stakes P_1 are strictly dominated by his betting rates \bar{P} , the bookmaker succeeds in ‘making a book’ which guarantees him a sure gain.

The often-repeated claim that ‘bookmakers must be Bayesians’ is mistaken: Bayesian bookmakers become bankrupt. Bookmakers aim to make a profit irrespective of the outcome ω . Bayesians cannot do that, because additive probabilities ‘avoid sure gain’ in the sense of section 2.8.2. A bookmaker who offered betting rates with the properties of additive probabilities would be unable to ‘make a book’, and he would not be a bookmaker (or a Bayesian) for long.

The same upper probability model \bar{P} is used also in life insurance. There $\bar{P}(A) = (1 + \delta)P_0(A)$ is the price of an insurance policy which pays out 1 unit if event A occurs, where now $P_0(A)$ is the actuarial probability of A .⁸

The conjugate pari-mutuel lower probability is

$$\underline{P}(A) = \max\{(1 + \delta)P_0(A) - \delta, 0\}.$$

As with the linear-vacuous model, $\bar{P}(A) - \underline{P}(A) = \delta$ is constant over events A for which $P_0(A)$ is not too close to 0 or 1 (i.e., $\tau \leq P_0(A) \leq 1 - \tau$). Unlike the linear-vacuous model, the pari-mutuel model satisfies $\bar{P}(A) \rightarrow 0$ as $P_0(A) \rightarrow 0$ and $\underline{P}(A) \rightarrow 1$ as $P_0(A) \rightarrow 1$.⁹

2.9.4 Capital gains tax and constant odds-ratio models

Suppose that when You purchase a risky investment, to be represented by a gamble X , any profit that You make on the investment is taxed at the constant rate τ , where $0 < \tau < 1$. When You pay price x for gamble X , Your net gain (after tax) on the investment is therefore $(1 - \tau)(X - x)^+ - (X - x)^-$. (Here $Z^+ = \max\{Z, 0\}$ and $Z^- = \max\{-Z, 0\}$ denote the positive and negative parts of Z , so that $Z = Z^+ - Z^-$.) If the linear prevision P_0 represents Your fair prices for gambles, the investment will be marginally desirable when $(1 - \tau)P_0((X - x)^+) = P_0((X - x)^-)$. Your supremum buying price for X , $\underline{P}(X)$, therefore satisfies the equation $(1 - \tau)P_0(G(X)^+) = P_0(G(X)^-)$, where as usual $G(X) = X - \underline{P}(X)$.

Define $f_X(x) = P_0((X - x)^-) - (1 - \tau)P_0((X - x)^+) = \tau P_0((X - x)^+) - P_0(X - x) = \tau P_0((X - x)^-) - (1 - \tau)P_0(X - x)$. Then f_X is a continuous, strictly increasing function (most easily seen from the last expression). Also

$f_X(\inf X) = -(1-\tau)(P_0(X) - \inf X) \leq 0$ and $f_X(P_0(X)) = \tau P_0((X - P_0(X))^+) \geq 0$. It follows that $\underline{P}(X)$ is the unique solution of the equation $f_X(x) = 0$, and satisfies $\inf X \leq \underline{P}(X) \leq P_0(X) \leq \sup X$. This defines, for all gambles X , a lower prevision $\underline{P}(X)$ which represents Your supremum buying price for gambles when profits are taxed at the constant rate τ . For reasons explained below, we call this lower prevision the **constant odds-ratio** (P_0, τ) model.

To see that \underline{P} is a coherent lower prevision, verify the axioms P1–P3. P1 holds as above. P2 holds because $(\lambda X - \lambda x)^+ = \lambda(X - x)^+$ when $\lambda > 0$. For P3, let $x = \underline{P}(X)$ and $y = \underline{P}(Y)$, so that $\tau P_0((X - x)^+) = P_0(X - x)$ and $\tau P_0((Y - y)^+) = P_0(Y - y)$. Using the fact that $(Z_1 + Z_2)^+ \leq Z_1^+ + Z_2^+$, we can obtain $\tau P_0((X + Y - x - y)^+) \leq \tau P_0((X - x)^+) + \tau P_0((Y - y)^+) = P_0(X + Y - x - y)$, so $f_{X+Y}(x + y) \leq 0$. Since f_{X+Y} is strictly increasing and $f_{X+Y}(\underline{P}(X + Y)) = 0$, it follows that $\underline{P}(X + Y) \geq x + y = \underline{P}(X) + \underline{P}(Y)$, so P3 holds. By Theorem 2.5.5, \underline{P} is coherent.

The conjugate upper prevision $\bar{P}(X)$ is the unique solution x of the analogous equation $P_0((X - x)^+) = (1 - \tau)P_0((X - x)^-)$, which is equivalent to $\tau P_0((X - x)^+) + (1 - \tau)P_0(X - x) = 0$.

These equations can be solved explicitly when $X = A$ is an event, since $(A - x)^+ = (1 - x)A$ when $0 \leq x \leq 1$. The resulting upper and lower probabilities are

$$\bar{P}(A) = \frac{P_0(A)}{1 - \tau P_0(A^c)} \quad \text{and} \quad \underline{P}(A) = \frac{(1 - \tau)P_0(A)}{1 - \tau P_0(A)}.$$

These satisfy $(1 - \tau)P_0(A) \leq \underline{P}(A) \leq P_0(A) \leq \bar{P}(A) \leq (1 - \tau)^{-1}P_0(A)$.

Now $\underline{P}(A)/\bar{P}(A^c) = (1 - \tau)P_0(A)/P_0(A^c)$ is the lower odds on A versus A^c , and $\bar{P}(A)/\underline{P}(A^c) = (1 - \tau)^{-1}P_0(A)/P_0(A^c)$ is the upper odds on A . The ratio of lower to upper odds on A is therefore $\underline{P}(A)\bar{P}(A^c)/\bar{P}(A)\underline{P}(A^c) = (1 - \tau)^2$, which is constant over all events A for which $0 < P_0(A) < 1$. (Hence the name ‘constant odds-ratio’).¹⁰

The constant odds-ratio (P_0, τ) lower probability dominates the linear-vacuous (P_0, τ) and pari-mutuel (P_0, τ) lower probabilities. That can be seen by interpreting the three lower probabilities as models for different kinds of taxation. Suppose You buy the gamble A for price x , where $0 \leq x \leq 1$. Under the constant odds-ratio model, Your profit on the transaction, $(A - x)^+ = (1 - x)A$, is taxed at rate τ . Under the linear-vacuous model, Your reward A is taxed at rate τ . Under the pari-mutuel model, Your maximum possible profit, $1 - x$, is taxed at rate τ .¹¹ Because Your profit cannot be larger than Your reward or Your maximum possible profit, i.e. $(1 - x)A \leq \min\{1 - x, A\}$, the constant odds-ratio tax is never larger than the other two taxes. In fact, the constant odds-ratio lower probability $\underline{P}(A)$ is strictly larger than the other two models unless $P_0(A) = 0$ or $P_0(A) = 1$. The ‘neighbourhood’ of P_0 consisting of those linear previsions that

dominate \underline{P} is therefore smaller for the constant odds-ratio model than for the other two models.¹²

The constant odds-ratio model is convenient for statistical applications because its form is not changed by conditioning or statistical updating. If the constant odds-ratio (P_0, τ) model is used to represent prior beliefs about a statistical parameter, and statistical data are obtained, then posterior beliefs are represented by the constant odds-ratio (P_1, τ) model, where P_1 is the posterior of P_0 defined by Bayes’ rule. See section 7.8.7 for details.¹³

2.9.5 ‘Uniform distributions’ on the positive integers

A uniform probability distribution on a finite space Ω assigns equal probability to each element of Ω . The next three examples are concerned with defining some kind of ‘uniform distribution’ when Ω is the set of positive integers \mathbb{Z}^+ , the unit interval $[0, 1]$, or the real line \mathbb{R} . In each case, non-linear lower previsions are natural models for a ‘uniform distribution’.

First consider $\Omega = \mathbb{Z}^+$. A ‘uniform distribution’ should assign the same probability to each positive integer, and it is clear that no countably additive probability model can do so. The usual approach is to construct a finitely additive model as a limit of uniform distributions on finite sets of integers.¹⁴ Let P_n denote the linear prevision generated by the uniform probability distribution on $\{1, 2, \dots, n\}$, so $P_n(X) = (1/n)\sum_{j=1}^n X(j)$ for $X \in \mathcal{L}(\mathbb{Z}^+)$. It is usual to define a finitely additive probability P by $P(A) = \lim_{n \rightarrow \infty} P_n(A)$, defined for just those subsets of \mathbb{Z}^+ for which the limit exists. However, this class of sets is unwieldy (it is not even a field).¹⁵ The probability P can be extended to a finitely additive probability on all subsets of \mathbb{Z}^+ , but not in any constructive way.

It is more natural to define a lower prevision \underline{P} on all gambles as the \liminf of the sequence $\{P_n : n \geq 1\}$, so $\underline{P}(X) = \liminf_{n \rightarrow \infty} (1/n)\sum_{j=1}^n X(j)$. This is a coherent lower prevision on $\mathcal{L}(\mathbb{Z}^+)$, by Corollary 2.6.7 (or verify axioms P1–P3, using super-linearity of \liminf). The conjugate upper prevision is $\bar{P}(X) = \limsup_{n \rightarrow \infty} (1/n)\sum_{j=1}^n X(j)$. We have $\underline{P}(X) = \bar{P}(X)$ just when $(1/n)\sum_{j=1}^n X(j)$ converges to a limit as $n \rightarrow \infty$. This happens for many events, e.g. $\underline{P}(A) = \bar{P}(A) = \frac{1}{2}$ when A is the set of even positive integers, but it is easy to construct sets A for which $\underline{P}(A) < \bar{P}(A)$, or even $\underline{P}(A) = 0$ and $\bar{P}(A) = 1$.

We will see in section 3.5 that \underline{P} is the lower envelope of a class of shift-invariant linear previsions (called ‘Banach limits’) on $\mathcal{L}(\mathbb{Z}^+)$. However, the lower envelope of the class of all shift-invariant linear previsions is strictly smaller than \underline{P} . It is the coherent lower prevision

$$\underline{Q}(X) = \liminf_{n \rightarrow \infty} \sum_{j=k}^{k+n-1} X(j),$$

which can be regarded as a lower limit of uniform probability distributions on finite sets $\{k, k+1, \dots, k+n-1\}$, as $n \rightarrow \infty$. Both \underline{P} and \underline{Q} are shift-invariant lower previsions.

For every finite set A , $(1/n) \sum_{j=1}^n A(j) \rightarrow 0$ so that $\underline{P}(A) = \bar{P}(A) = 0$, and similarly $\underline{Q}(A) = \bar{Q}(A) = 0$. Any additive probability P that dominates \underline{P} or \underline{Q} is dominated by \bar{P} or \bar{Q} and so satisfies $P(A) = 0$ for all finite sets A , which is incompatible with countable additivity of P .¹⁶

Although the models \underline{P} and \underline{Q} are coherent, we cannot automatically conclude that they are reasonable models. We will see in Chapter 6 that they lack a further rationality property of ‘full conglomerability’: for some choices of conditioning events, there are no conditional previsions that are coherent with these models. (The same is true of all their dominating linear previsions.)¹⁷

2.9.6 Inner and outer Lebesgue measure

Consider next the case where $\Omega = [0, 1]$, the unit interval. We require a ‘uniform distribution’ on Ω to be **translation-invariant**. That is, if we define a translation operator \oplus for $x, y \in \Omega$ by $x \oplus y = x + y$ if $x + y < 1$, $x \oplus y = x + y - 1$ if $x + y \geq 1$, and write $B \oplus y = \{x \oplus y : x \in B\}$ for the translate of a set B , we require that the translates $B \oplus y$ have the same probability for all $y \in \Omega$. It is well known that Lebesgue measure, v , is the unique translation-invariant, countably additive probability measure on the Borel σ -field \mathcal{B} of Ω .¹⁸ Here \mathcal{B} is defined as the smallest σ -field containing all intervals $[\alpha, \beta]$ (equivalently, all open subsets) of Ω , and v is defined as the unique countably additive probability on \mathcal{B} such that the probability of each interval is equal to its length, $v([\alpha, \beta]) = \beta - \alpha$.

Lebesgue measure has no translation-invariant, countably additive extension to all subsets of $[0, 1]$.¹⁹ However, it does have a natural extension to translation-invariant upper and lower probabilities defined on all subsets of $[0, 1]$. These are the **outer and inner Lebesgue measure**, defined by

$$\bar{P}(A) = \inf \{v(B) : B \supseteq A, B \in \mathcal{B}\} \quad \text{and} \quad \underline{P}(A) = \sup \{v(B) : A \subseteq B, B \in \mathcal{B}\}$$

for all subsets A .²⁰ The **Lebesgue measurable** sets A are those for which $\underline{P}(A) = \bar{P}(A)$.²¹ At the other extreme, there are non-measurable sets A for which $\underline{P}(A) = 0$ and $\bar{P}(A) = 1$.²²

One way to see that inner and outer Lebesgue measure are coherent is to extend them to all gambles. Let \mathcal{K} denote the class of all \mathcal{B} -measurable gambles, and let P_0 be the unique linear prevision on \mathcal{K} that extends Lebesgue measure. Define the **Lebesgue lower prevision** on $\mathcal{L}(\Omega)$ by $\underline{P}(X) = \sup \{P_0(Y) : Y \leq X, Y \in \mathcal{K}\}$. First note that this lower prevision does

agree on events with the inner measure: it clearly dominates inner measure, and it is dominated by inner measure because whenever $Y \in \mathcal{K}$ with $Y \leq A$, we can define $B = \{\omega \in \Omega : Y(\omega) > 0\}$ so that $B \in \mathcal{B}$, $A \supseteq B$, and $Y \leq B \leq A$. Similarly, outer Lebesgue measure is the restriction to events of the conjugate upper prevision $\bar{P}(X) = \inf \{P_0(Y) : Y \geq X, Y \in \mathcal{K}\}$. Next verify that the Lebesgue lower prevision satisfies axioms P1–P3. It is therefore coherent, by Theorem 2.5.5. Hence inner Lebesgue measure, its restriction to events, is a coherent lower probability.

In fact, as shown in section 3.4, inner and outer Lebesgue measure are the lower and upper envelopes of the class of all finitely additive extensions of Lebesgue measure to all subsets of Ω . The smaller class of finitely additive extensions that are translation-invariant is characterized in section 3.5, in terms of its upper and lower envelopes.

2.9.7 ‘Uniform distributions’ on the real line

Now let $\Omega = \mathbb{R}$, the real line. Again, a ‘uniform distribution’ on \mathbb{R} should be translation-invariant. Writing $B + y = \{x + y : x \in B\}$, this means that the translates $B + y$ should have the same probability for all $y \in \mathbb{R}$. There is no countably additive probability measure that is translation-invariant, even on the intervals. (To see that, note that \mathbb{R} is the countable union of intervals $[n, n+1]$ over integers n , and these intervals have the same probability, which must be zero by finite additivity.)

Analogously to the example of the positive integers (section 2.9.5), we can define translation-invariant upper and lower previsions as limits of uniform distributions on finite intervals. Let \mathcal{K} be the linear space of Borel-measurable gambles on \mathbb{R} , and define $\underline{P}(X) = \liminf_{c \rightarrow \infty} (1/2c) \int_{-c}^c X(t) dt$. This is a lower limit of uniform distributions on intervals centred at zero, as their length $2c \rightarrow \infty$. It is a coherent lower prevision on \mathcal{K} , by Corollary 2.6.7. It can be extended to a translation-invariant, coherent lower prevision on $\mathcal{L}(\mathbb{R})$, by defining $\underline{P}(Y) = \sup \{\underline{P}(X) : X \leq Y, X \in \mathcal{K}\}$.

Again by analogy with section 2.9.5, a second translation-invariant lower prevision can be defined on \mathcal{K} by $\bar{Q}'(X) = \lim_{c \rightarrow \infty} \inf_{x \in \mathbb{R}} (1/2c) \int_{x-c}^{x+c} X(t) dt$. This is a lower limit of uniform distributions on intervals $[x - c, x + c]$ whose length $2c \rightarrow \infty$. Clearly \underline{P} dominates \bar{Q}' , and \bar{Q}' can be much less precise than \underline{P} . When $A = \mathbb{R}^+$ (the set of positive reals), for example, $\underline{P}(A) = \bar{P}(A) = \frac{1}{2}$, but $\bar{Q}'(A) = 0$, $\bar{Q}'(A) = 1$. (Here \bar{P} and \bar{Q}' are the conjugate upper previsions, obtained by replacing ‘inf’ by ‘sup’ in the definitions of \underline{P} and \bar{Q}' .) In fact, there are Borel sets A (countable unions of intervals) with $\underline{P}(A) = 1$ but $\bar{Q}'(A) = 0$. Since $\bar{P}(A) = \bar{Q}'(A) = 0$ for all bounded sets A , none of the additive probabilities dominated by \bar{P} or \bar{Q}' can be countably additive.

The model \bar{Q}' appears naturally in the statistical problems discussed later,

as a kind of coherent replacement for the improper uniform prior on \mathbb{R} . (See especially Example 8.2.9.)²³

2.9.8 Filters and zero-one valued probabilities

As a final example, consider the lower probabilities which take only the values 0 and 1. We show that there is a one-to-one correspondence between coherent 0–1 valued lower probabilities and non-empty filters of sets. A class \mathcal{A} of subsets of Ω is called a **filter** when it satisfies

- (a) $\emptyset \notin \mathcal{A}$
- (b) $A \cap B \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$
- (c) $B \in \mathcal{A}$ whenever $\Omega \supset B \supset A$ and $A \in \mathcal{A}$.

A non-empty filter contains Ω , by (c), and cannot contain both A and A^c , by (a) and (b).

Given any coherent 0–1 valued lower probability \underline{P} defined on all subsets of Ω , let $\mathcal{A} = \{A : \underline{P}(A) = 1\}$ be the collection of events which are almost certain under \underline{P} . By results 2.7.4, coherence implies the properties $\underline{P}(\emptyset) = 0$, $\underline{P}(A) + \underline{P}(B) \leq 1 + \underline{P}(A \cap B)$, and $\underline{P}(B) \geq \underline{P}(A)$ whenever $B \supset A$, which imply (a), (b) and (c) respectively in the definition of a filter. Since $\underline{P}(\Omega) = 1$, \mathcal{A} is a non-empty filter.

Given any non-empty filter \mathcal{A} , define a 0–1 valued lower probability by $\underline{P}(A) = 1$ if $A \in \mathcal{A}$, $\underline{P}(A) = 0$ otherwise. To verify that \underline{P} is coherent, use the general definition 2.5.1.²⁴ Note that $G(A) = A - 1 \leq 0$ when $A \in \mathcal{A}$, and $G(A) = A \geq 0$ when $A \notin \mathcal{A}$. It therefore suffices to prove that $\sup[\sum_{j=1}^n (A_j - 1) - mA_0] \geq 0$ whenever $A_0 \notin \mathcal{A}$ and $A_1, \dots, A_n \in \mathcal{A}$. This holds because $B = \bigcap_{j=1}^n A_j \in \mathcal{A}$ by property (b) of filters, $B \cap A_0^c$ is non-empty by property (c), and $\sum_{j=1}^n (A_j(\omega) - 1) - mA_0(\omega) = 0$ for any $\omega \in B \cap A_0^c$. Thus \underline{P} is a coherent lower probability. This establishes the one-to-one correspondence between filters and coherent 0–1 valued lower probabilities.

The conjugate upper probability for the filter \mathcal{A} is also 0–1 valued, given by $\bar{P}(A) = 0$ if $A^c \in \mathcal{A}$, $\bar{P}(A) = 1$ otherwise. The degree of imprecision concerning A is $\bar{P}(A) - \underline{P}(A) = 0$ if either $A \in \mathcal{A}$ or $A^c \in \mathcal{A}$, and $\bar{P}(A) - \underline{P}(A) = 1$ otherwise. So the 0–1 valued lower probabilities admit only the three extreme kinds of judgements about an event A : that A is almost certain to occur ($A \in \mathcal{A}$), that A is almost certain not to occur ($A^c \in \mathcal{A}$), or that there is complete ignorance concerning A (neither $A \in \mathcal{A}$ nor $A^c \in \mathcal{A}$).

When Ω is finite, all filters have the form $\mathcal{A} = \{A : A \supset A_0\}$ for some non-empty A_0 , and all coherent 0–1 valued lower probabilities have the form $\underline{P}(A) = 1$ if $A \supset A_0$, $\underline{P}(A) = 0$ otherwise. When Ω is infinite, there are other kinds of filters and 0–1 valued lower probabilities.²⁵

The vacuous lower probability (2.9.1) corresponds to the smallest

non-empty filter $\mathcal{A} = \{\Omega\}$. At the other extreme, additive probabilities correspond to maximal filters, which are called **ultrafilters**. A filter \mathcal{A} is an ultrafilter if it satisfies the maximality property:

- (d) for every event A , either $A \in \mathcal{A}$ or $A^c \in \mathcal{A}$.

The property of coherent 0–1 valued lower probabilities which corresponds to (d) is the self-conjugacy property $\underline{P}(A) + \underline{P}(A^c) = 1$, that is, $\underline{P}(A) = \bar{P}(A)$ for all events A . By Lemma 2.8.8, self-conjugacy characterizes the additive probabilities. Thus there is a one-to-one correspondence between ultrafilters and 0–1 valued additive probabilities.

Extending a filter to an ultrafilter corresponds to finding a 0–1 valued additive probability that dominates a 0–1 valued lower probability. Expressing a filter as an intersection of ultrafilters corresponds to expressing a coherent 0–1 valued lower probability as a lower envelope of 0–1 valued additive probabilities. That can be done using the results of section 3.6.

The natural extension of a coherent 0–1 valued lower probability \underline{P} to a lower prevision (defined for all gambles X) is

$$\underline{E}(X) = \sup \{\mu : \underline{P}([X \geq \mu]) = 1\} = \sup \{\mu : [X \geq \mu] \in \mathcal{A}\},$$

where $[X \geq \mu]$ denotes the event $\{\omega : X(\omega) \geq \mu\}$ and \mathcal{A} is the filter corresponding to \underline{P} . (This will be proved in section 3.2.) Thus $\underline{E}(X)$ is a kind of ‘essential infimum’ for X ; it is the supremum value of μ for which You are almost certain that $X \geq \mu$. To see that this defines a coherent lower prevision on $\mathcal{L}(\Omega)$, verify axioms P1–P3 using the properties of a filter. When \mathcal{A} is an ultrafilter, \underline{E} is a linear prevision.

As an example,²⁶ let $\Omega = \mathbb{Z}^+$ and let \mathcal{A} be the filter consisting of all subsets of Ω whose complements are finite. If A is any finite set then $\bar{P}(A) = 0$; You are almost certain that A will not occur. The natural extension of this model to a lower prevision is $\underline{E}(X) = \sup \{\mu : [X < \mu] \text{ is finite}\} = \liminf_{n \rightarrow \infty} X(n)$, with conjugate upper prevision $\bar{E}(X) = \limsup_{n \rightarrow \infty} X(n)$. (The ‘essential infimum’ of X is $\liminf X$ rather than $\inf X$, as You assign zero probability to any finite set of integers.) This model is less precise than the ‘uniform’ models \underline{P} and \bar{Q} defined in section 2.9.5. In fact, $\bar{E} \geq \bar{Q} \geq \bar{P} \geq \underline{P} \geq \underline{Q} \geq \underline{E}$.

2.10 Interpretations of lower previsions

Now that the basic properties of coherent upper and lower previsions have been presented and illustrated through examples, it is time for a closer look at how previsions and probabilities are to be interpreted. On the minimal behavioural interpretation of lower previsions (section 2.3.1), \underline{P} represents a disposition to accept gambles of the form $X - \mu$ whenever $\mu < \underline{P}(X)$. In this section we shall discuss some of the ways we might extend this

behavioural interpretation. The various types of extension will be called aleatory, logical, personalist, operational, exhaustive and sensitivity analysis interpretations. The mathematical theory of coherence is consistent with this wide range of extended interpretations, but we do not believe that any of these extensions is generally applicable to problems of reasoning under uncertainty. The behavioural interpretation that we favour, which might be characterized (in contrast to the extended interpretations) as epistemic, rationalistic, theoretical, incomplete and direct, does seem to be generally applicable. Of course, the extended interpretations are applicable in some specific kinds of problems.

2.10.1 Aleatory versus epistemic

The behavioural interpretation outlined in section 2.3.1 may seem to have little in common with aleatory (frequentist or propensity) interpretations of probability. But most frequentists do believe that aleatory probabilities, when known, can be used in decision making to determine which actions are reasonable. Indeed many have advocated some kind of principle of direct inference: when You know the values of aleatory probabilities, You should adopt them as rational epistemic probabilities. Hence the coherence conditions, and the mathematical theory based on them, have some force even for aleatory probabilities.

Amongst epistemic interpretations there is a basic distinction between logical and personalist interpretations. Our behavioural interpretation is consistent with both of these, although neither seems adequate in general. We sometimes do have compelling reasons for making probability judgements (e.g., when we know aleatory probabilities, or when we have absolutely no relevant evidence), and then a logical interpretation is applicable. But there are usually great difficulties in establishing a particular probability model as uniquely rational, and it is implausible that this can be done in problems where the available evidence is complicated or ambiguous.

We are somewhat less sympathetic to personalist interpretations as rationality clearly involves more than coherence. We prefer a rationalistic interpretation which requires probability assessments to be justified by reference to the evidence and assessment strategies on which they are based, while allowing that there may be several reasonable assessment strategies which lead to different probability models. (See section 1.7 for discussion.)

2.10.2 Operational versus theoretical¹

Operational interpretations take P to represent specific gambling commitments, to accept on demand any gamble of the form $X - \mu$ where

$\mu < P(X)$. This requires You to act as a bookmaker by posting a list of betting rates (or buying or selling prices) at which You are willing to accept bets. To give an unambiguous operational definition, it is necessary to specify exactly what the commitments involve (restrictions on the stakes or the number of gambles that will be accepted, to whom the gambles are offered, for what period of time the commitments apply, etc.).

There are applications in which an operational interpretation is appropriate. When \bar{P} represents the betting rates set by bookmakers (as in section 2.9.3 and Appendix C), it would be wrong to regard \bar{P} as a model for ‘underlying beliefs’ and an operational interpretation seems necessary. Similarly, participants in the experiment described in Appendix I were told that by specifying upper and lower probabilities they were committing themselves to accept particular gambles.

The advantages of an operational interpretation are that it enables probabilities to be elicited in a straightforward way through the corresponding operational measurement procedure (Appendix H), and it provides a more immediate justification for avoiding sure loss. If the rewards of gambles are expressed in some ‘public’ currency such as money or probability currency, and Your gambling commitments incur sure loss, then it is likely that You will be punished by someone who wishes to make a sure gain.

However, operational interpretations have serious disadvantages which make them unsuitable for most applications. In particular, (a) it is unclear why operationally elicited probabilities should characterize Your behaviour in contexts different from that of elicitation, (b) ‘beliefs’ are identified with specific behaviour rather than an underlying state of mind, (c) probability models with an operational interpretation cannot be falsified or corrected, (d) assertions about the desirability of hypothetical gambles cannot be admitted as evidence about beliefs, and (e) beliefs about unobservable states cannot be given an operational interpretation.

Theoretical interpretations do distinguish specific actions from underlying beliefs, by interpreting P as a model for Your dispositions to act in a wide range of hypothetical circumstances. Part of the meaning of the model P is that You will accept gambles of the form $\lambda(X - \mu)$ where $\lambda > 0$ and $\mu < P(X)$, provided the rewards are in units of linear utility. But the model describes Your state of mind (disposition to accept such gambles) rather than any specific gambling commitments, and has implications for Your behaviour in more complicated decision problems when utilities are imprecise. Thus P is regarded as a theoretical model of a real psychological state (Your beliefs) that influences observable behaviour (Your actions and assertions) but is not itself directly observable.

2.10.3 Exhaustive versus incomplete

Suppose $\underline{P}(X)$ and $\bar{P}(X)$ are defined for some gamble X , with $\underline{P}(X) < \bar{P}(X)$. We interpret this model as asserting Your willingness to accept gambles of the form $X - \mu$ for $\mu < \underline{P}(X)$ and $\mu - X$ for $\mu > \bar{P}(X)$. If it is assumed that Your dispositions are coherent, it can be inferred also that You are unwilling to accept $X - \mu$ for $\mu > \bar{P}(X)$ or $\mu - X$ for $\mu < \underline{P}(X)$. On our interpretation, however, nothing can be said about Your attitude to $X - \mu$ or $\mu - X$ when $\underline{P}(X) \leq \mu \leq \bar{P}(X)$. You may be entirely willing to accept such a gamble, or entirely unwilling to do so, or have various degrees of willingness, or be wholly undecided. Under this interpretation, \underline{P} may be an incomplete description of Your beliefs.²

In contrast, under an exhaustive interpretation something definite can be said about Your attitude to $X - \mu$ when $\underline{P}(X) < \mu < \bar{P}(X)$. In that case You are undecided about whether to accept $X - \mu$. You have no disposition to accept it, but nor do You have a disposition to reject it. Then $\underline{P}(X)$ is uniquely defined as the supremum price μ for which You have a disposition to accept $X - \mu$, $\bar{P}(X)$ is the infimum price μ for which You have a disposition to reject $X - \mu$, and this characterizes Your attitude to gambles $X - \mu$ for all real μ (except for the two borderline values $\mu = \underline{P}(X)$ and $\mu = \bar{P}(X)$). So \underline{P} is an exhaustive description of Your dispositions concerning gambles in its domain.³

Under an exhaustive interpretation, the only source of imprecision in a model \underline{P} is real indeterminacy in beliefs. Our interpretation allows extra imprecision due to incomplete elicitation of beliefs. An exhaustive model, when it can be obtained, is more useful than an incomplete model because it is more informative about beliefs.

We believe that exhaustive models can sometimes be constructed, but they will often be very difficult to elicit, intractable or unnecessary. For example, a model \underline{P} which is constructed through a careful analysis of evidence may be adopted as an exhaustive model. That is, when the model fails to determine a unique action You will not try to resolve the indecision by further elicitation or analysis of evidence. An exhaustive interpretation is also reasonable when the evidence can be completely specified and there is a unique assessment strategy for translating it into a model \underline{P} , so that \underline{P} can be given a logical interpretation.

There are various reasons why the models \underline{P} used in practice may fail to be exhaustive.

- Elicitation is likely to be incomplete due to limited time and effort. In the elicitation procedure studied in Chapter 4, $\underline{P}(X)$ is the supremum price for X that is implied by Your explicit assessments, which need not be exhaustive.

- It may be unnecessary to make probability assessments as precise as they can be for events that are of minor importance.
- Previsions can be automatically extended to a larger domain by natural extension (section 3.1), but they may not be an exhaustive model for beliefs about the larger domain.
- An intractable model \underline{P} may be replaced by a less precise but more tractable model \underline{Q} that is dominated by \underline{P} . Then \underline{Q} will be an incomplete model for beliefs even if \underline{P} is exhaustive. This can be especially useful when \underline{Q} is a conjugate prior or has properties of independence.⁴
- If the members of a group have beliefs modelled by $\underline{P}_1, \dots, \underline{P}_n$, then the group beliefs can be modelled by the lower envelope $\underline{P}(X) = \min\{\underline{P}_j(X) : 1 \leq j \leq n\}$. The model \underline{P} can then be regarded as an incomplete description of the beliefs of each member of the group.
- Under an exhaustive interpretation, there is a sharp boundary between the gambles You are disposed to accept and those You are not disposed to accept. Elicitation of an exhaustive \underline{P} therefore requires the same kind of precise discriminations as the Bayesian theory.⁵ To construct an incomplete model, it suffices to identify some gambles that You are definitely disposed to accept. The remaining gambles will include some You are definitely not disposed to accept, and others towards which Your attitude is ambiguous.⁶

For these reasons we do not require that models \underline{P} be exhaustive, although they will often be exhaustive in simple problems where the evidence is easy to specify and assess. In practice we will usually try to make \underline{P} as exhaustive (precise) as possible, subject to the above qualifications, through careful assessment and elicitation.

2.10.4 Direct versus sensitivity analysis

So far we have considered **direct** interpretations of the lower prevision \underline{P} , which take \underline{P} to directly describe Your behavioural dispositions, without reference to the linear previsions of which it is lower envelope. **Sensitivity analysis** interpretations, on the other hand, regard \underline{P} merely as a convenient summary of a class of linear previsions. According to the dogma of ideal precision, Your beliefs should (at least as an ideal) be represented by some linear prevision P_T , and a non-linear model \underline{P} is needed in practice only because of uncertainty about the correct P_T .

Suppose, as an example, that You judge an event A to be ‘probable’. On our behavioural interpretation, this means just that You are willing to bet on A at even stakes. The sensitivity analysis interpretation is that A has some ‘ideal’ precise probability which You judge to be greater than $\frac{1}{2}$.

The sensitivity analysis interpretation is implicit in the practice of Bayesian sensitivity analysis. It is also the interpretation adopted by most previous students of imprecise probabilities (see section 1.8). Its popularity may be partly due to the results of Smith (1961) and Williams (1976), which show that every coherent lower prevision is the lower envelope of a class of linear previsions (see section 3.3).

Under sensitivity analysis interpretations, imprecision of the model \underline{P} arises solely from incomplete assessment or elicitation, and not from indeterminacy in ideal beliefs. Compare with exhaustive interpretations, which admit indeterminacy but not incompleteness as a source of imprecision. (See section 5.2 for discussion of the two sources.) Our interpretation admits both sources and is therefore consistent with both sensitivity analysis and exhaustive interpretations. But we see less justification for a sensitivity analysis interpretation than for an exhaustive one. In principle, models can be made exhaustive by sufficiently thorough analysis of the available evidence, whereas indeterminacy is more fundamental and can be eliminated only by obtaining more evidence.

A sensitivity analysis interpretation is reasonable in those problems where You have a sufficiently large amount of information to justify a precise probability model, but You lack the time, computational ability or assessment strategies to completely analyse this information. You may then consider various linear previsions as hypotheses about the ideal beliefs that You would obtain from a complete analysis. In our view, there is rarely sufficient information to justify precise probabilities even as an ideal. (See Chapter 5 for further discussion.)

There are several versions of the sensitivity analysis interpretation which differ according to the meaning ascribed to the ‘true’ or ‘ideal’ linear prevision P_T . There is surprisingly little discussion of the meaning of P_T in the large literature on sensitivity analysis, but we can distinguish the following interpretations.

(a) Descriptive

If interpreted descriptively, P_T represents Your real (unconscious) behavioural dispositions, about which You have only partial information. That is, Your real beliefs are coherent and determinate in the ways needed for representation by a linear prevision P_T (e.g. Your selling prices for gambles agree with buying prices), and the imprecise model \underline{P} arises from incomplete elicitation or ‘inexact measurement’ of P_T .⁷ But, in the light of psychological studies of human reasoning and behaviour, as well as naïve introspection, it is wholly implausible that behavioural dispositions can be accurately represented by precise probabilities, especially when the dispositions are unconscious.⁸ Most Bayesians recognize that a descriptive interpretation is untenable and prefer some normative version.

(b) Logical

One normative interpretation is that P_T represents the beliefs that would result from an ideal analysis of the available evidence, if only You had the time and ability to carry this out. Then P_T can be given a logical interpretation, as the uniquely rational model on this evidence.⁹ In practice Your analysis of the evidence may be incomplete, again leading to imprecision in the model \underline{P} . The questionable assumption here is that a complete analysis would justify precise logical probabilities. It is argued in Chapter 5 that, whereas precision may be reasonable when there is sufficient evidence, probabilities should be imprecise when they are based on little information.

(c) Personalist

Personalist Bayesians cannot use the previous interpretation because they cannot say how probabilities should be ‘ideally’ assessed from evidence. A personalist version is that P_T represents the beliefs that You would eventually settle on, if You had unlimited time and resources to use for assessment.¹⁰ The imprecision in the model \underline{P} arises from Your uncertainty about the beliefs You would settle on. But it is again assumed that You would be able to settle on determinate beliefs, represented by a linear prevision P_T . That is unclear, as personalists cannot say how P_T should be constructed. On what grounds should You choose one linear prevision that dominates \underline{P} rather than another?

Our conclusion is that sensitivity analysis interpretations involve very strong and implausible assumptions about the precision of beliefs. They rely on the dogma of ideal precision, that You should aim at the ‘ideal’ of precise probabilities. If this dogma is rejected, it is not clear what meaning can be given to the linear previsions P that dominate an elicited lower prevision \underline{P} . In this book we eschew all sensitivity analysis interpretations and restrict ourselves to the direct behavioural interpretation of lower previsions.¹¹

2.10.5 What difference does it make?

It may appear that, because a lower prevision can always be given a sensitivity analysis interpretation in addition to its behavioural interpretation, the difference in interpretation will not matter in practice. That is true only up to a point. Most of the theory presented in this book can be regarded as a formalization of Bayesian sensitivity analysis, although we would not advocate that interpretation. That includes all of the theory of unconditional previsions (Chapters 2–4), and much of the theory of conditioning and statistical models (Chapters 6–8). There is a mathematical correspondence between coherent lower previsions and classes of linear previsions

(Theorems 3.3.3, 3.6.1), and this extends to conditional probabilities and statistical models provided these are defined on finite spaces (Appendix K3). The correspondence does not extend to infinite spaces under our definitions of coherence. For example, there are conglomerable lower previsions that are not lower envelopes of conglomerable linear previsions (section 6.9.8).

More serious differences between the two interpretations appear when we define structural properties of independence and permutability (Chapter 9). Our behavioural interpretation leads us to consider two events as independent when betting rates on either one conditional on the other are the same as unconditional betting rates. (So knowledge that one event has occurred is irrelevant when betting on the other.) On the other hand, a sensitivity analyst will regard the events as independent when they are independent under all the additive probabilities in the class being considered. The two definitions are quite different, and the difference may be important in practice because it will affect how we translate intuitive judgements that events are ‘unrelated’ into a mathematical model.

In decision problems, the two approaches are mathematically equivalent provided that utility assessments are precise and the class of feasible actions is closed under randomization. But differences arise when both probabilities and utilities are imprecise. Then sensitivity analysts would restrict attention to the actions that are optimal with respect to some pair of precise probability and precise utility functions, whereas the behavioural interpretation suggests that other actions can be reasonable (Example 3.9.10).

Perhaps the most important practical effect of interpretation is on the assessment and elicitation of probability models. It is clear that the operational, exhaustive and sensitivity analysis interpretations can lead to quite different probability assessments from ours, because the imprecision of probabilities has a different meaning, and arises from different sources, under each interpretation. In statistical problems where there is little prior information, we would adopt highly imprecise prior probabilities, whereas sensitivity analysts would presumably adopt a precise (‘noninformative’) prior distribution. More generally, sensitivity analysts would construct a model by considering properties of the ‘ideal’ precise probabilities, whereas we would consider behavioural properties of the imprecise model.¹²

The interpretation of lower previsions affects the development of the mathematical theory because it suggests definitions of important concepts such as maximality (in decision making), conditional previsions, independence, permutability, and extended notions of coherence for conditional previsions and statistical models. All these concepts will be defined here so as to have a direct behavioural interpretation, without reference to dominating linear previsions.

2.11 Objections to behavioural theories of probability

In this section we discuss some common criticisms of behavioural theories of probability. These are possible objections to Bayesian theories as well as the theory presented here, though we believe that the present theory is better able to answer them. Other criticisms that are essentially concerned with precision (the over-precision of Bayesian models, or the imprecision of upper and lower previsions) will be deferred to Chapter 5. We ignore the many criticisms of behavioural theories that are simply due to misunderstandings. The objections listed here deserve to be taken seriously, especially those concerning prior beliefs (2.11.1), objectivity (2.11.2), external rationality (2.11.7) and probability assessment (section 2.11.8), which are closely related. Despite these objections, we believe that a behavioural interpretation of probability provides the soundest foundation for theories of statistical inference and decision. Our discussion of the objections concentrates on statistical inference because that is the main topic of this book, and also because a behavioural interpretation is less controversial in decision making than in inference.¹

2.11.1 *The emphasis on belief and behaviour*

A common objection is that science should not be contaminated by personal beliefs. Statistical inference, in particular, should not be contaminated by prior beliefs, which are seen as a source of bias or prejudice. A different objection is that inference (belief) should be clearly distinguished from decision (behaviour). The emphasis in our approach on beliefs and behaviour, and the close connection between them, may therefore seem misguided. We would reply to these objections along the following lines.

1. The conclusions of inference are epistemic; they reflect a particular state of information. Statistical reasoning is reasoning about an external reality (aleatory probabilities), but it is reasoning nevertheless, and therefore inherently epistemic. The conclusions of statistical reasoning are usually uncertain. Some kind of epistemic probability is needed to express the uncertainty about statistical hypotheses that is warranted by the available information.
2. These epistemic probabilities should have some kind of behavioural interpretation, so that statistical conclusions will be potentially applicable as guides to future inquiry and decision. Here we do not require inferences to be aimed at some specific decision; we accept that inference is quite different from decision. Nor do we require that epistemic probabilities be understood merely as models for behavioural dispositions. But it seems reasonable to require that statistical conclusions can be used as a basis

for action, even if no particular action is currently envisaged. That is guaranteed by the behavioural interpretation adopted here. The role in inquiry or decision of other measures of statistical uncertainty, such as confidence coefficients, observed significance levels or likelihood ratios, is not so clear.

3. The aim of statistical inference is to draw conclusions from statistical data. The epistemic probabilities used to measure the uncertainty in these conclusions should be conditional on the observed statistical data, since it is the uncertainty after seeing the data that is relevant. That is, the relevant probabilities are ‘posterior’ probabilities. This, as well as various arguments in favour of the likelihood principle, counts against frequentist methods of statistical inference. (See section 7.5 for the case of confidence intervals.)

4. Typically, statistical data alone generate vacuous posterior probabilities for statistical hypotheses. To achieve useful (non-vacuous) posterior probabilities, some non-vacuous prior probabilities are needed. (This argument is given in section 7.4.1.) In effect, this means that we must consider ‘prior information’, meaning whatever information about the statistical hypotheses is available apart from the statistical data. Another argument for this is that conclusions and decisions should be based on all the available information, not just the statistical data. If there is substantial prior information then it would be foolish to ignore it. These arguments count against statistical methods that use only the observed likelihood function.

5. Despite our emphasis on ‘beliefs’, we do not advocate a personalist interpretation of epistemic probabilities. If a statistical inference is to have interpersonal validity, the prior probabilities on which it is based must be more than just models for personal beliefs. Instead they must be justified as ‘externally rational’ assessments of prior information. The posterior probabilities obtained from statistical inference then represent ‘rational beliefs’. Personal judgements are necessary in statistical inference, especially in choosing models and assessment strategies, but the judgements should be recognized and supported, as far as possible, by referring to the evidence on which they are based.²

6. Unless there is very substantial prior information, precise prior probabilities cannot be justified. When prior probabilities are imprecise, posterior probabilities are too, and their precision increases with the amount of information in the statistical data. When there is little prior information, or when prior assessments are not available, the prior probabilities should be highly imprecise, and in some sense ‘close to’ the vacuous prior. (See the Bernoulli examples in section 5.3 for one possible sense.) Statistical conclusions will be interpersonally acceptable to all reasonable statisticians,

and even to personalist Bayesians, provided the prior probabilities are sufficiently imprecise.

The preceding argument (all of whose steps are controversial) suggests to us that statistical conclusions should be expressed in terms of imprecise posterior probabilities which are based on all the available evidence (statistical data plus prior information), and whose interpretation is epistemic, behavioural, and rationalistic or logical.

2.11.2 *Objectivity*

The most common objections to behavioural theories, and to many people the strongest objections, concern the lack of ‘objectivity’ of epistemic probabilities, especially the prior probabilities used in Bayesian statistics. Several different issues have been confused. To disentangle them it is necessary to distinguish several different kinds of objectivity, which will be called (a) physicality, (b) interpersonal agreement, (c) external rationality, and (d) well-groundedness.³

Many Bayesians have regarded objectivity as reducible to **interpersonal agreement**, or consensus, amongst qualified observers.⁴ For example, there is often greater agreement amongst scientists over how statistical data were generated (i.e., over a sampling model or likelihood function) than over prior probabilities of hypotheses. On this view, the greater ‘objectivity’ of the sampling model is merely a reflection of the greater agreement.

Most critics of the Bayesian theory have advocated a different kind of objectivity, which we will call **physicality**.⁵ A probability model is objective in this sense when its correctness depends only on its correspondence with external reality, and not on the knowledge or beliefs of any observers. This involves an aleatory interpretation of probability, in which probability models refer to the outside world, whereas interpersonal agreement involves only an epistemic interpretation (agreement amongst the beliefs of various observers). Obviously, interpersonal agreement about a model is neither necessary nor sufficient for it to correspond to physical reality.

We believe that the business of science is to formulate and test hypothetical explanations for real phenomena. The probabilities that appear in scientific theories that make plausible claims about reality (e.g. quantum mechanics or genetics) should therefore be given an aleatory interpretation.⁶ It follows that there is reason to distinguish statistical sampling models involving aleatory hypotheses from prior distributions involving merely epistemic probabilities, but it does not follow that epistemic priors are useless or inadmissible in statistical inference. This position is contrary to Bayesians like de Finetti who rule out aleatory probabilities, and to frequentists like

Neyman who admit prior probabilities only when they have a frequentist interpretation.

As a simple example, consider tossing a thumbtack. The usual Bernoulli model is that successive tosses are physically independent with some constant chance θ of landing pin-up. Here θ is regarded as a physical property of the tack and tossing mechanism. Information about the value of θ can be obtained from repeated tosses and reflected in posterior epistemic probabilities for θ , as illustrated in section 5.3.

One can also interpret the ‘objectivity’ of this model in terms of interpersonal agreement. Suppose several people each assess precise personal probabilities for unlimited sequences of outcomes of future tosses. If each person regards the future tosses as exchangeable then their probabilities can be represented, through de Finetti’s theorem, in terms of the Bernoulli model together with a personal prior distribution for θ . (See section 9.5 for details.) The Bernoulli model is ‘objective’ in the sense that it abstracts the features of the personal beliefs about which there is interpersonal agreement. As a large number of tosses are observed, there is, under fairly general conditions, increasing agreement about the true value of θ . ‘The true value’ means, on this account, the value to which personal probabilities of pin-up on a future toss would converge with increasing data. Thus the purported ‘objectivity’ of the aleatory model is reduced to interpersonal agreement.⁷

Unfortunately, such a reduction loses the most important feature of the aleatory model – its physicality. On the reductionist account, a group of people who agreed on exchangeability could not learn that the Bernoulli model was incorrect (e.g. because of correlation between successive outcomes). The only sense of ‘correctness’ admitted by the reductionist account is that of long-run interpersonal agreement, and a group initially agreed on exchangeability will, if they update their beliefs in a coherent way, remain agreed on exchangeability. Their model may be incorrect in the physicalist sense, despite the interpersonal agreement, because it may not correspond to the real aleatory probabilities.

Even if the Bernoulli model is physically correct, there need not be convergence of opinions with increasing data, e.g. when one person is sure that $\theta \leq \frac{1}{3}$ and another is sure that $\theta \geq \frac{2}{3}$. In that case we surely want to say that at least one of them is wrong. Or, if all members of the group are sure that $\theta \leq \frac{1}{3}$ and yet relative frequencies of pin-up converge to $\frac{2}{3}$, then there may be convergence of opinions concerning θ , but increasing certainty that θ is close to $\frac{1}{3}$! It seems perverse to regard $\frac{1}{3}$ as the ‘true value’ in that case. Even if these cases are ruled out by requiring that all prior distributions for θ have support $[0, 1]$, so that posterior distributions for θ are increasingly concentrated near the limit of relative frequencies, it is surely more natural to regard this limit as a reflection of physical reality rather than of the

group’s beliefs. The increasing interpersonal agreement can be explained most naturally by admitting aleatory probabilities as hypotheses about the thumbtack, and recognizing that the group is learning about the correct probabilities.

We recognize, of course, that much scientific activity is concerned with identifying, clarifying and reducing interpersonal disagreements between scientific experts. When experts assess different probability distributions P_j , for example, it is important to identify the areas of agreement and disagreement. A simple way to do so is to regard the lower envelope of expert beliefs, $\underline{P}(X) = \min\{P_j(X): 1 \leq j \leq n\}$, as a model for group beliefs. Then the imprecision of the group beliefs concerning an event A , $\Delta(A) = \bar{P}(A) - \underline{P}(A)$, directly measures the extent of disagreement amongst group members concerning the probability of A . If A concerns the value of a statistical parameter θ , as in the thumbtack example, the imprecision $\Delta(A)$ will typically be reduced as statistical evidence accumulates.

The third and fourth meanings of objectivity concern the relation of probabilities to evidence. Under the logical interpretations of Keynes, Carnap and Jeffreys, probabilities are epistemic but also ‘objective’, in the sense that they are uniquely determined by logical or linguistic relationships between hypothesis and evidence. Without demanding uniqueness we might call epistemic probabilities more or less objective to the extent that they are **externally rational** (see section 1.7). Epistemic probabilities are said to be **well grounded** when they are based on a large body of relevant information.

Both external rationality and well-groundedness are distinct from physicality, since they concern epistemic rather than aleatory probabilities, and also from interpersonal agreement, since they concern the relation between probability and evidence. However, well-grounded probabilities are likely to be interpersonally acceptable to those who share the evidence on which they are based. Thus sampling models based on extensive frequency data are likely to be interpersonally acceptable. Epistemic probabilities based on less extensive information are more likely to be interpersonally acceptable when they are derived using well-established assessment strategies, i.e. when they can be shown to be externally rational.

2.11.3 Different types of uncertainty

A related criticism is that behavioural theories of probability treat all types of uncertainty alike, because all probabilities are interpreted in terms of behavioural dispositions. It is often argued that uncertainty about the outcome of a physically random experiment should be distinguished from uncertainty about the value of an unknown physical constant.⁸ The two

cases can, in fact, be distinguished in several ways, using several senses of ‘objectivity’.

First, sampling models for the outcome of a random experiment usually have an aleatory interpretation, whereas epistemic probabilities concerning the value of a statistical parameter usually do not. This difference in interpretation is not necessarily reflected in different numerical probabilities.

A second distinction is that statistical sampling models are often based on extensive statistical data or physical understanding of a phenomenon (they are well grounded), whereas prior distributions are often based on little evidence. We distinguish these two cases through precision of probabilities.⁹ Well grounded sampling models will be precise, whereas prior distributions based on little evidence will be highly imprecise. Bayesians, on the other hand, represent both kinds of information by precise probabilities. Frequentist and likelihood theories of statistics do distinguish the two kinds of information, but they are explicitly concerned only with statistical information. There are no formal methods for incorporating non-statistical prior information even when it is substantial.¹⁰ Both Bayesian and frequentist approaches seem unsatisfactory.

Similarly, qualitative probability judgements can be distinguished from precise numerical ones, or the assessments of experts from those of laymen, through their precision. These can be viewed as ‘different types’ of uncertainty, but they can be covered, while recognizing the distinctions, by a single behavioural interpretation and a unified theory of coherence.

2.11.4 *The emphasis on gambling*

Some people find behavioural theories objectionable because of their emphasis on gambles, betting rates and avoiding sure loss.¹¹ Specific objections are:

1. thinking about betting rates (or buying prices for gambles) is often not a good way to assess probabilities (or previsions);
2. the ‘gambles’ to which the behavioural interpretation refers are highly artificial, as they involve rewards of probability currency or other linear utility;
3. the betting situations considered are ‘unfair’ to You, since You are forced to offer betting rates whereas an opponent is free to choose his bets and stakes.¹²

We do not disagree with the first assertion but, we do not see it as an objection. It is often difficult to directly assess reasonable buying and selling prices for gambles. Other methods of constructing probability models may be more useful. (Many of these are outlined in Chapter 4.) But that is quite

consistent with our behavioural interpretation. We require only that probability models, however they are constructed, have implications for buying prices, betting rates and other behaviour.

As for objection 2, gambles are indeed artificial constructions. The artificiality is deliberate, to simplify the interpretation and measurement of previsions by controlling the influence of utilities on choices. A characteristic feature of gambles is that the utilities of their rewards are precise. In practical decision problems, the utilities of rewards are typically imprecise. Your beliefs can be most easily understood through their effect on Your attitudes to gambles, but they also influence Your behaviour in other decision problems.

Objection 3 is not a valid criticism of the present theory. You are not forced to offer any betting rates at all. Nor need there be any ‘opponent’ who aims to exploit Your rates.

2.11.5 *The emphasis on avoiding sure loss*

To satisfy the basic rationality criterion, avoiding sure loss, there must be no finite combination of desirable gambles that is certain to produce an overall loss. It is distasteful to many people that they should be so concerned with ‘protecting themselves’ against sure loss, especially in problems of inference where rewards and losses may seem irrelevant. Specific objections are:

1. Avoiding sure loss has force only when a ‘sure loss’ can be exploited by an opponent; that cannot happen in statistical inference, nor even in most decision problems.
2. When Your probability assessments incur sure loss, it is highly unlikely in practice that You will be confronted with the specific combination of gambles that produces the sure loss.
3. Even if that unlikely event happens, You can avoid harm simply by refusing to accept some of the gambles whose combination incurs sure loss.¹³

Our reply to 1 and 2 is that if Your previsions incur sure loss then You are disposed to act, in some hypothetical and admittedly artificial situations, in ways which would certainly produce an undesirable result. That seems to indicate fundamental irrationality in Your assessments, irrespective of whether the situations actually arise and You actually do lose. The force of avoiding sure loss comes not so much from the fear of actual loss as from the realization that Your probability assessments would lead You, in these hypothetical situations, to willingly take actions whose net result is sure to be harmful. The criterion seems just as compelling in problems of

inference, where no specific action is envisaged, as in decision problems. Of course, in decision problems the previsions that incur sure loss may lead immediately to decisions that are obviously bad, especially when several choices are made simultaneously or when the decision depends on the outcome of a statistical experiment.¹⁴

Those who remain disturbed by the emphasis on avoiding sure loss may prefer to regard coherence (a stronger property) as the fundamental criterion. Coherence can be regarded as a kind of self-consistency, without reference to ‘losses’. (That is especially clear from the axioms in section 2.3.3.) In particular, coherence may seem more relevant than avoiding sure loss to problems of statistical inference.

As for 3, if You intend to ‘avoid sure loss’ in practice by refusing to accept some of the gambles that You are supposedly ‘disposed’ to accept, then Your lower previsions do not represent invariable dispositions to accept gambles. Rather, there must be some interaction between separate choices; the desirability of a gamble X depends on what other gambles have already been chosen. That is tantamount to a violation of axiom D3: You are disposed to accept gamble X and also disposed to accept gamble Y , but not disposed to accept one of X and Y selected according to the toss of a coin. (See section 2.2.4.) If Your dispositions to accept gambles can interact in this way then it becomes very complicated to model them and to ensure that they do avoid sure loss.

2.11.6 Coherence is too strong

Some objections of this sort are that:

1. psychological studies of human reasoning and decision have shown that the coherence condition is often violated;
2. people cannot be fully coherent because of their limited abilities in computation and in processing information;
3. we should not expect people to satisfy coherence until they are shown how to construct coherent probability models which properly reflect their information;
4. other rationality criteria are more important than coherence, and it is sometimes necessary to violate coherence in order to satisfy them;
5. in applications of the theory, incoherence is less important than other sources of error, such as those arising from the idealizations and approximations needed to construct tractable probability models.

The extent to which coherence is violated in practice is not clear. Most psychological studies have been restricted to artificial problems in which there are correct aleatory probabilities to be assessed, and have assumed that subjects satisfied the Bayesian axioms of precision.¹⁵ Consistent violations of the Bayesian coherence axioms have been observed in some

experiments, e.g. intransitivity of choices and conservatism in updating probabilities.¹⁶ The Bayesian axioms do not appear to be a good description of real beliefs, but it seems that at least some of the violations may be due to the elicitation of precise probabilities from subjects whose beliefs are actually imprecise, e.g. forced choices may be intransitive when they do not reflect genuine preferences. Imprecise probability models seem to be better descriptions of real beliefs than the Bayesian models. However, coherence of probability assessments is unlikely to be achieved without explicit analysis. (The weaker conditions of n -coherence, defined in Appendix B, are easier to satisfy.) We view the present theory, as most Bayesians view their theory, as a means of improving and extending human reasoning abilities, rather than as a description of naïve human reasoning.

There clearly are computational and assessment problems of the sort suggested in 2 and 3. The need to develop useful assessment strategies, 3, seems to be the most fundamental of these problems. It is addressed in the next two subsections. We regard coherence not as an obstacle to assessment and information processing, but rather as an essential guide in carrying out these tasks.

Our attitude to 4 is similar. We agree that there are other important rationality criteria, but we do not know of any that are incompatible with coherence.¹⁷ For example, it is more important that probabilities are soundly based on evidence (or externally rational) than that they are exactly coherent. However, since we regard coherence and natural extension as central in the process of constructing probabilities from evidence, we do not see any fundamental incompatibility between coherence and external rationality.

Concerning 5, the need for idealizations and approximations in constructing probability models is no reason, in itself, to add a further source of error by making incoherent inferences from these models! However, there are problems in which exact coherence may have to be sacrificed, in order to reduce other sources of error or to increase tractability. For example, it may not be feasible to compute exact inferences, by natural extension, from a complicated model. It may be better to build a complex and realistic statistical model, from which only approximate inferences can be drawn, than to build a simplistic model that allows exact inferences. As in any branch of applied mathematics, it may often be necessary to depart from an ideal (coherent) analysis because of practical constraints. But it is important to recognize such departures as sources of error, and to estimate their effect on conclusions.

2.11.7 Coherence is too weak

We agree that coherence is too weak to fully characterize ‘rationality’. Further criteria of external rationality are needed to ensure that probability

models conform to the available evidence. Such criteria should include the four principles described in section 1.7.2. (See section 1.7 for further discussion of this issue.)

2.11.8 Probability assessment

A related objection to behavioural theories is that they give little guidance on how to sensibly analyse evidence in order to assess probabilities. This is a very serious objection which can be answered only by developing a repertoire of assessment strategies and models that apply to particular kinds of evidence. Many simple assessment strategies will be described in Chapters 4 to 9. Obviously much more work is needed to develop these into an adequate repertoire. In our view, the assessment problem can be handled only when imprecise assessments are admitted, as these are needed to properly reflect a lack of information or conflicting information.

2.11.9 Observability

On our behavioural interpretation, probability models describe dispositions to accept gambles defined on Ω . An objection to this is that it applies only when the true state ω can be determined, so that the values of gambles become known. That seems to rule out cases where ω is not observable, notably those where ω parametrizes a statistical sampling model. As our account of statistical inference is concerned with beliefs about statistical parameters, we need to show that such beliefs are meaningful.¹⁸

Call a possibility space Ω **observable** if there is a specific ‘verification procedure’ that can be applied, at or before some specified time, to determine a single element of Ω called the ‘true state’.¹⁹ (Derivatively, a state or event is observable if it can be determined whether or not it is or contains the true state.) The outcomes of random experiments such as coin-tossing are usually observable, whereas the values of statistical parameters are usually unobservable. If θ denotes the chance that a particular coin has of landing ‘heads’, then the value of θ , or even the event ($\theta > \frac{1}{2}$) that the coin is biased towards ‘heads’, is unobservable. Bets on this event cannot be conclusively settled in any finite time.

When Ω is unobservable, gambles on Ω may never be settled and lower previsions cannot be given the behavioural interpretation outlined in section 2.3.1. They can still be given a behavioural interpretation but it must be indirect. Beliefs about an unobservable space must be capable of influencing behaviour in at least some hypothetical situations. In particular, beliefs about a statistical parameter space Θ influence Your beliefs about future statistical observations. Your beliefs concerning Θ can therefore be

interpreted and measured indirectly, through Your attitudes to gambles defined on observables whose probability distribution depends on θ .²⁰

As a simple example, let $\Theta = \{0.4, 0.6\}$ represent two hypotheses about the bias of a coin, and $\mathcal{X} = \{H, T\}$ the two possible outcomes of a single toss. Suppose Your lower and upper probabilities for the hypothesis that $\theta = 0.6$ are α and β . These values cannot be directly elicited, but Your upper and lower probabilities for ‘heads’ on the toss, $\bar{P}(H)$ and $\underline{P}(H)$, can be elicited from the rates at which You are willing to bet on ‘heads’. The coherence conditions in Chapter 6 imply the relations $\underline{P}(H) = 0.4 + 0.2\alpha$ and $\bar{P}(H) = 0.4 + 0.2\beta$, so that α and β can be obtained from $\alpha = 5\underline{P}(H) - 2$ and $\beta = 5\bar{P}(H) - 2$. In this case, Your betting rates fully reveal Your beliefs about Θ .

In most cases, beliefs about Θ will not be fully determined by beliefs about just a few observables. In the coin-tossing example, if we extend Θ to the usual probability interval $[0, 1]$, then assessments of $\underline{P}(H)$ and $\bar{P}(H)$ provide only partial information about beliefs concerning Θ . When A is the event that $\theta > \frac{1}{2}$, they imply the imprecise assessments $\underline{P}(A) = \max\{2\underline{P}(H) - 1, 0\}$ and $\bar{P}(A) = \min\{2\bar{P}(H), 1\}$, which are vacuous when $\underline{P}(H) \leq \frac{1}{2} \leq \bar{P}(H)$. More information can be obtained by eliciting Your upper and lower probabilities concerning sequences of independent tosses.

In many cases it may be possible to accurately elicit Your beliefs about an unobservable space Ω by simply asking what Your attitudes to gambles defined on Ω would be, under the assumption that ω could be determined so that the gambles could be settled. As the hypothetical gambles could not be settled, there are several potential difficulties. It may be difficult to get You to take such hypothetical gambles seriously, and there may be a psychological bias in favour of states ω which, if true, are most likely to be observable. The hypothetical nature of the gambles does not seem an overwhelming objection. Even when elicitation concerns observables, it will usually involve hypothetical gambles whose rewards are not actually paid. Your motivation for taking the elicitation process seriously is that Your assessments will influence the inferences or decisions You make.

This approach is most useful when some sort of ‘approximate verification procedure’ is conceivable. The chance θ that a coin lands ‘heads’ is not observable (in a finite time), but it can be estimated, with arbitrarily high probability of attaining arbitrarily high accuracy, from the outcomes of sufficiently many independent tosses. Some gambles concerning θ , such as bets on the event that $\theta > \frac{1}{2}$, can be settled with very high probability of correctness on the basis of a long sequence of observations. Gambles may then be evaluated on the assumption that they will be settled by such an ‘approximate verification procedure’.

CHAPTER 3

Extensions, envelopes and decisions

In this chapter we develop the mathematical theory of coherence by introducing some alternative models for uncertainty and studying their relation to lower previsions. The three main topics of the chapter are:

1. extension of a lower prevision to larger domains, especially through natural extension (sections 3.1 and 3.2 – the results are applied to linear previsions in sections 3.4 and 3.5);
2. the relationship between a lower prevision and its class of dominating linear previsions (sections 3.3–3.6);
3. models for desirability and preference, their relation to lower previsions, and their application in decision making (sections 3.7–3.9).

In Chapter 2 we modelled beliefs through a lower prevision $(\Omega, \mathcal{K}, \underline{P})$. No restrictions were placed on the possibility space Ω or domain \mathcal{K} . This generality is useful as it enables us to construct \underline{P} on whatever domain is most convenient. To develop the model it may be necessary to extend \underline{P} to a larger domain. In section 3.1 we show that coherent lower previsions always have a coherent extension to any larger domain. The minimal coherent extension to the class of all gambles, which is called the natural extension and denoted by \underline{E} , represents the behavioural implications of \underline{P} . The concept of natural extension has an important role in the ensuing theory. Section 3.2 deals with the problem of extending a coherent lower probability \underline{P} from a field \mathcal{A} to the class of \mathcal{A} -measurable gambles. When \underline{P} is additive on \mathcal{A} , its natural extension agrees with the usual definition of linear expectation. More generally, if \underline{P} has a property of 2-monotonicity then its natural extension can be computed by integrating upper distribution functions.

In section 3.3 we examine the relationship between a lower prevision \underline{P} and its class of dominating linear previsions $\mathcal{M}(\underline{P})$. The separating hyperplane theorem is used to show that \underline{P} avoids sure loss if and only if $\mathcal{M}(\underline{P})$ is non-empty, and that \underline{P} is coherent on domain \mathcal{K} if and only if it is the lower envelope of $\mathcal{M}(\underline{P})$ on \mathcal{K} . In general, the lower envelope of $\mathcal{M}(\underline{P})$ on \mathcal{L} , the class of all gambles, is the natural extension of \underline{P} . This result is

3.1 NATURAL EXTENSION

applied in section 3.4 to characterize the linear extensions of a given linear prevision. (These always exist.) It is used again in section 3.5 to characterize the linear previsions that are invariant under a semigroup of transformations. Invariant linear previsions exist, and there is a simple formula for their lower envelope, provided the semigroup is Abelian. These results are used to characterize the Banach limits on the positive integers and the translation-invariant extensions of Lebesgue measure.

Topological ideas are introduced in section 3.6 to establish a one-to-one correspondence between coherent lower previsions \underline{P} on \mathcal{L} and compact convex sets of linear previsions, and to show that each coherent \underline{P} is the lower envelope of the extreme points of $\mathcal{M}(\underline{P})$. Application of these results to the example of zero–one valued lower probabilities yields the ultrafilter theorem and establishes the existence of non-degenerate ultrafilters and non-measurable sets.

Sections 3.7 to 3.9 are concerned with models for desirability and preference. We distinguish almost-desirability from strict desirability. (A more general notion of desirability is discussed in Appendix F.) Each type of desirability corresponds to a type of preference: gamble X is preferred to Y if and only if $X - Y$ is desirable. Axioms are given in section 3.7 to characterize coherence of the four models. There are natural one-to-one correspondences, defined in section 3.8, between coherent lower previsions and coherent models for the four types of desirability and preference. For example, the coherent classes \mathcal{D} of almost-desirable gambles are just the closed convex cones containing all non-negative gambles but not all gambles, and \mathcal{D} corresponds to \underline{P} by $\underline{P}(X) = \max\{\mu: X - \mu \in \mathcal{D}\}$ and $\mathcal{D} = \{X: \underline{P}(X) \geq 0\}$. The merits of these alternative models are discussed in section 3.8.6.

A theory of decision making is outlined in section 3.9. We assume that each possible action is evaluated through a linear utility function, which can be regarded as a gamble. The reasonable actions are those whose gambles are maximal under \underline{P} , i.e., there is no better gamble in the strict partial preference ordering corresponding to \underline{P} . Our main result (3.9.5) is that, provided randomized decisions are allowed, maximality under \underline{P} is equivalent to maximality under a linear prevision that dominates \underline{P} . An example is given to show this does not hold when both utilities and probabilities are imprecise.

3.1 Natural extension

After the ideas of avoiding sure loss and coherence studied in the previous chapter, the most important concept in the present theory is that of natural extension. It is the fundamental concept in our theory of statistical inference (Chapter 8) and in our approach to elicitation (Chapter 4). Indeed, natural

extension may be seen as the basic constructive step in statistical reasoning; it enables us to construct new previsions from old ones.

In fact, the three key ideas (avoiding sure loss, coherence and natural extension) are very closely connected. Given a lower prevision \underline{P} that is defined on domain \mathcal{K} and avoids sure loss, there is a minimal coherent lower prevision \underline{E} , defined on all gambles, that dominates \underline{P} on \mathcal{K} . When \underline{P} is coherent on \mathcal{K} , \underline{E} is the minimal coherent extension of \underline{P} to all gambles.¹ This \underline{E} will be called the **natural extension** of \underline{P} . It summarizes the buying prices for gambles that are implied by \underline{P} through the linear operations involved in the definition of coherence. For every gamble X , $\underline{E}(X)$ is the largest buying price for X that can be constructed from Your specified buying prices $\underline{P}(X_j)$ through linear operations.

The definition of natural extension is already implicit in our general definition of coherence (2.5.1). From the equivalent condition 2.5.4(a), incoherence occurs just when there are gambles X_0, X_1, \dots, X_n in \mathcal{K} , $\lambda_j \geq 0$ and $\alpha > \underline{P}(X_0)$ such that $X_0 - \alpha \geq \sum_{j=1}^n \lambda_j G(X_j)$. The gamble on the right of the inequality should be at least marginally desirable, and hence You should be willing to pay up to α for X_0 . Since α exceeds the specified supremum buying price $\underline{P}(X_0)$, there is incoherence. More generally, this argument shows that in order to achieve coherence Your lower prevision $\underline{P}(X_0)$ must be at least as large as the supremum value of α for which there are $X_j \in \mathcal{K}$ and $\lambda_j \geq 0$ such that $X_0 - \alpha \geq \sum_{j=1}^n \lambda_j G(X_j)$. It turns out that this value of $\underline{P}(X_0)$ does achieve coherence, and we take it as our definition of natural extension.

3.1.1 Definition of natural extension

Suppose that $(\Omega, \mathcal{K}, \underline{P})$ is a lower prevision. Then \underline{E} , called the **natural extension** of \underline{P} , is defined on domain $\mathcal{L} = \mathcal{L}(\Omega)$ by

$$\underline{E}(X) = \sup \left\{ \alpha : X - \alpha \geq \sum_{j=1}^n \lambda_j G(X_j) \text{ for some } n \geq 0, X_j \in \mathcal{K}, \lambda_j \geq 0, \alpha \in \mathbb{R} \right\}.$$

As a simple example of natural extension, suppose that You assess upper and lower probabilities for two events A and B that are *logically independent*, meaning that $A \cap B$, $A \cap B^c$, $A^c \cap B$ and $A^c \cap B^c$ are all non-empty. Suppose that Your assessments satisfy the coherence constraints $0 \leq \underline{P}(A) \leq \bar{P}(A) \leq 1$ and $0 \leq \underline{P}(B) \leq \bar{P}(B) \leq 1$. What do they imply about the probabilities of $A \cup B$ and $A \cap B$? It can be verified, using Theorems 3.1.2(f), 2.7.4(a, c, d, f) and 3.4.1, that the natural extensions of the four assessments are

$$\underline{E}(A \cup B) = \max \{\underline{P}(A), \underline{P}(B)\}, \bar{E}(A \cup B) = \min \{1, \bar{P}(A) + \bar{P}(B)\},$$

$$\underline{E}(A \cap B) = \max \{0, \underline{P}(A) + \underline{P}(B) - 1\}, \text{ and } \bar{E}(A \cap B) = \min \{\bar{P}(A), \bar{P}(B)\}.$$

3.1 NATURAL EXTENSION

These new probabilities are quite imprecise. Even if precise probabilities are assessed for A and B , they are quite uninformative about the probabilities of $A \cup B$ and $A \cap B$.

The natural extension of \underline{P} is trivial unless \underline{P} avoids sure loss. If \underline{P} incurs sure loss then $\sum_{j=1}^n \lambda_j G(X_j)$ can be made arbitrarily large by taking λ_j to be sufficiently large, hence α in the definition can be made arbitrarily large and $\underline{E}(X) = \infty$ for every gamble X . On the other hand, if \underline{P} avoids sure loss and $X - \alpha \geq \sum_{j=1}^n \lambda_j G(X_j)$ then $\sup X - \alpha \geq \sup \sum_{j=1}^n \lambda_j G(X_j) \geq 0$, hence $\underline{E}(X) \leq \sup X$. Note also that $\underline{E}(X) \geq \inf X$, by taking $n = 0$ and $\alpha = \inf X$ in the definition. Thus \underline{P} avoids sure loss if and only if its natural extension \underline{E} is finite.

Coherence of \underline{P} can also be characterized in terms of the natural extension. The following theorem shows that \underline{P} is coherent on domain \mathcal{K} if and only if its natural extension \underline{E} agrees with \underline{P} on \mathcal{K} . In that case \underline{E} is a genuine ‘extension’ of \underline{P} . When \underline{P} avoids sure loss but is not coherent, its natural extension \underline{E} dominates \underline{P} on \mathcal{K} , but $\underline{E}(X)$ must disagree with $\underline{P}(X)$ for some X in \mathcal{K} . (In fact, $\underline{E}(X_0) > \underline{P}(X_0)$ for just those $X_0 \in \mathcal{K}$ for which the coherence condition 2.5.1 is violated.) In that case \underline{E} ‘corrects’ \underline{P} on \mathcal{K} and is not (strictly speaking) an ‘extension’ of \underline{P} . Thus the natural extension functions both to make \underline{P} coherent on its domain \mathcal{K} , and to extend \underline{P} beyond \mathcal{K} in a coherent way.

3.1.2 Basic properties of natural extension

Suppose the lower prevision $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss. Then \underline{E} , its natural extension to \mathcal{L} , has the following properties:

- (a) $\inf X \leq \underline{E}(X) \leq \sup X$ for all $X \in \mathcal{L}$
- (b) \underline{E} is a coherent lower prevision on \mathcal{L}
- (c) \underline{E} dominates \underline{P} on \mathcal{K} , i.e. $\underline{E}(X) \geq \underline{P}(X)$ for all $X \in \mathcal{K}$
- (d) \underline{E} agrees with \underline{P} on \mathcal{K} if and only if \underline{P} is coherent
- (e) \underline{E} is the minimal coherent lower prevision on \mathcal{L} that dominates \underline{P} on \mathcal{K}
- (f) if \underline{P} is coherent then \underline{E} is the minimal coherent extension of \underline{P} to \mathcal{L} .

Proof.

- (a) was proved earlier.
- (b) Verify axioms P1–P3 of section 2.3.3: P1 holds by (a), and P2 and P3 follow from definition of \underline{E} .
- (c) Take $n = 1, X_1 = X$ and $\alpha = \underline{P}(X)$ in the definition.
- (d) Suppose \underline{P} is coherent and $X \in \mathcal{K}$. If $X - \alpha \geq \sum_{j=1}^n \lambda_j G(X_j)$ then $\underline{P}(X) - \alpha \geq \sup [\sum_{j=1}^n \lambda_j G(X_j) - G(X)] \geq 0$ by 2.5.4, so that $\alpha \leq \underline{P}(X)$. Hence $\underline{E}(X) \leq \underline{P}(X)$, and (c) gives $\underline{E}(X) = \underline{P}(X)$. Conversely, if \underline{E} agrees with \underline{P} on \mathcal{K} then \underline{P} is coherent by (b).
- (e) holds by a similar argument to (d). Suppose \underline{Q} is coherent on \mathcal{L} and \underline{Q}

dominates \underline{P} on \mathcal{K} . If $X - \alpha \geq \sum_{j=1}^n \lambda_j G(X_j)$ then $\alpha \leq \inf[X - \sum_{j=1}^n \lambda_j (X_j - \underline{P}(X_j))] \leq \inf[X - \sum_{j=1}^n \lambda_j (X_j - Q(X_j))] \leq Q(X)$, using coherence of Q . Hence $\underline{E}(X) \leq Q(X)$ for all $X \in \mathcal{L}$. Thus \underline{E} is minimal. (f) follows from (d) and (e). ◆

In view of (e) and (f), the natural extension might also be called the ‘minimal’ or ‘least-committal’ extension. It makes the minimal claims about behavioural dispositions that are consistent with \underline{P} .

By (f), any coherent lower prevision, defined on an arbitrary domain, does have a coherent extension to a lower prevision defined on all gambles. This is an elementary consequence of the definition of coherence, whereas the corresponding result for linear previsions (Theorem 3.4.2) requires more advanced mathematical tools (the separating hyperplane theorem, or equivalent). Moreover, it is often not possible to explicitly specify a linear extension of a linear prevision, whereas the natural extension is always defined by 3.1.1.

The natural extension \underline{E} is not generally the unique coherent extension of \underline{P} , even when \underline{P} is defined on a ‘large’ domain such as the class of all events, as illustrated by Example 2.7.3. But any other coherent extension of \underline{P} must strictly dominate \underline{E} and therefore models behavioural dispositions that are not implied by \underline{P} . Necessary and sufficient conditions for \underline{E} to be the unique coherent extension are given in Theorem 3.1.7.

The natural extension can be written in the following alternative forms.

3.1.3 Lemma²

Suppose $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss. The natural extension \underline{E} and its conjugate upper prevision \bar{E} are given (for all $X \in \mathcal{L}$) by the formulas:

- (a) $\underline{E}(X) = \sup \left\{ \inf \left[X - \sum_{j=1}^n \lambda_j G(X_j) \right]: n \geq 0, X_j \in \mathcal{K}, \lambda_j \geq 0 \right\}$
- (b) $\underline{E}(X) = \sup \left\{ \sum_{j=1}^n \lambda_j \underline{P}(X_j) + \mu: X \geq \sum_{j=1}^n \lambda_j X_j + \mu, n \geq 0, X_j \in \mathcal{K}, \lambda_j \geq 0, \mu \in \mathbb{R} \right\}$
- (c) $\bar{E}(X) = \inf \left\{ \beta: \beta - X \geq \sum_{j=1}^n \lambda_j G(X_j), n \geq 0, X_j \in \mathcal{K}, \lambda_j \geq 0, \beta \in \mathbb{R} \right\}$
- (d) $\bar{E}(X) = \inf \left\{ \sup \left[X + \sum_{j=1}^n \lambda_j G(X_j) \right]: n \geq 0, X_j \in \mathcal{K}, \lambda_j \geq 0 \right\}$
- (e) $\bar{E}(X) = \inf \left\{ \sum_{j=1}^n \lambda_j \bar{P}(X_j) + \mu: X \leq \sum_{j=1}^n \lambda_j X_j + \mu, n \geq 0, X_j \in -\mathcal{K}, \lambda_j \geq 0, \mu \in \mathbb{R} \right\}$

The formulas for natural extension simplify when the domain \mathcal{K} of \underline{P} has suitable structure. The following results treat the cases where \mathcal{K} is a convex cone of gambles or a field of events.

3.1 NATURAL EXTENSION

3.1.4 Theorem

Suppose \underline{P} is a coherent lower prevision defined on a convex cone \mathcal{K} . Then its natural extension \underline{E} is given by

$$\begin{aligned} \underline{E}(X) &= \sup \{ \alpha: X - \alpha \geq G(Y), Y \in \mathcal{K} \} \\ &= \sup \{ \underline{P}(Y) + \mu: X \geq Y + \mu, Y \in \mathcal{K}, \mu \in \mathbb{R} \}, \end{aligned}$$

and the conjugate upper prevision is

$$\begin{aligned} \bar{E}(X) &= \inf \{ \beta: \beta - X \geq G(Y), Y \in \mathcal{K} \} \\ &= \inf \{ \bar{P}(Y) + \mu: X \leq Y + \mu, Y \in -\mathcal{K}, \mu \in \mathbb{R} \}. \end{aligned}$$

If also \mathcal{K} contains all constant gambles then these formulas simplify to

$$\underline{E}(X) = \sup \{ \underline{P}(Y): X \geq Y, Y \in \mathcal{K} \} \quad \text{and} \quad \bar{E}(X) = \inf \{ \bar{P}(Y): X \leq Y, Y \in -\mathcal{K} \}.$$

Proof. The first formula follows from Definition 3.1.1 by writing $Y = \sum_{j=1}^n \lambda_j X_j$, since $Y \in \mathcal{K}$ and $G(Y) \leq \sum_{j=1}^n \lambda_j G(X_j)$. The second formula holds by writing $\mu = \alpha - \underline{P}(Y)$, and the formulas for \bar{E} by writing $\bar{E}(X) = -\underline{E}(-X)$ and $\beta = -\alpha$. If also \mathcal{K} contains all constant gambles, then $Y + \mu \in \mathcal{K}$ and $\underline{P}(Y + \mu) = \underline{P}(Y) + \mu$. ◆

3.1.5 Theorem

Suppose \underline{P} is a coherent lower probability defined on a field \mathcal{A} . Then its natural extension to a coherent lower probability defined on all events is $\underline{E}(B) = \sup \{ \underline{P}(A): B \supset A, A \in \mathcal{A} \}$. The conjugate upper probability is $\bar{E}(B) = \inf \{ \bar{P}(A): A \supset B, A \in \mathcal{A} \}$. (Extension of \underline{P} to other gambles is discussed in section 3.2.)

Proof. Suppose $B - \alpha \geq \sum_{j=1}^n \lambda_j G(A_j)$, where $A_j \in \mathcal{A}$ and $\lambda_j \geq 0$. Write $Y = \sum_{j=1}^n \lambda_j G(A_j) + \alpha$, so $B \geq Y$. Let $A = \{\omega: Y(\omega) > 0\}$. Then $A \in \mathcal{A}$, since \mathcal{A} is a field. When $Y(\omega) > 0$, $A(\omega) = 1 = B(\omega) \geq Y(\omega)$. When $Y(\omega) \leq 0$, $B(\omega) \geq 0 = A(\omega) \geq Y(\omega)$. Thus $B \geq A \geq Y$, so $B \supset A$. Using coherence of \underline{P} , $0 \leq \sup[\sum_{j=1}^n \lambda_j G(A_j) - G(A)] = \sup[Y - \alpha - A + \underline{P}(A)] \leq \underline{P}(A) - \alpha$, so that $\alpha \leq \underline{P}(A)$. So

$$\begin{aligned} \underline{E}(B) &= \sup \left\{ \alpha: B - \alpha \geq \sum_{j=1}^n \lambda_j G(A_j) \quad \text{for } A_j \in \mathcal{A}, \lambda_j \geq 0 \right\} \\ &\leq \sup \{ \underline{P}(A): B \supset A, A \in \mathcal{A} \}. \end{aligned}$$

The reverse inequality holds by taking $n = 1, A_1 = A, \alpha = \underline{P}(A)$. ◆

When \mathcal{A} is a σ -field (closed under countable unions and intersections), the supremum in the theorem is attained by some $A \in \mathcal{A}$, so $\underline{E}(B) = \max \{ \underline{P}(A): B \supset A, A \in \mathcal{A} \}$ and $\bar{E}(B) = \min \{ \bar{P}(A): A \supset B, A \in \mathcal{A} \}$. (To see that, let $A_n \in \mathcal{A}, B \supset A_n$, $\underline{E}(B) \leq \underline{P}(A_n) + n^{-1}$ and take $A = (\bigcup_{n=1}^{\infty} A_n)$. When P is a

countably additive probability defined on a σ -field, its natural extension \underline{E} to all subsets of Ω is called the **inner measure** generated by P , and the conjugate upper probability \bar{E} is the **outer measure** generated by P .³

3.1.6 Range of coherent extensions

Suppose \underline{P} is a coherent lower prevision defined on domain \mathcal{K} . Consider its coherent extensions \underline{Q} to \mathcal{L} . We wish to characterize the range of values $\underline{Q}(Z)$ for an arbitrary gamble Z in $\mathcal{L} - \mathcal{K}$. We need only consider coherent extension to $\mathcal{K} \cup \{Z\}$ since any such extension can be coherently extended to all of \mathcal{L} . The only constraint on the value of $\underline{Q}(Z)$ is that it must be coherent with the values of \underline{P} on \mathcal{K} . Applying the coherence condition 2.5.4(a), we obtain the constraints

$$\inf \left[Z - \sum_{j=1}^n \lambda_j G(X_j) \right] \leq \underline{Q}(Z) \leq \sup \left[Z + \sum_{j=1}^n \lambda_j G(X_j) - \lambda_0 G(X_0) \right]$$

for all $X_j \in \mathcal{K}$ and $\lambda_j \geq 0$.

By Lemma 3.1.3(a), the supremum of the lower bound is the natural extension $\underline{E}(Z)$, and this is the minimal value of $\underline{Q}(Z)$ as \underline{Q} ranges over coherent extensions. The maximal coherent value of $\underline{Q}(Z)$ is

$$\begin{aligned} \underline{E}(Z) &= \inf \left\{ \sup \left[Z + \sum_{j=1}^n \lambda_j G(X_j) - \lambda_0 G(X_0) \right] : X_j \in \mathcal{K}, \lambda_j \geq 0 \right\} \\ &= \inf \left\{ \beta : \beta - Z \geq \sum_{j=1}^n \lambda_j G(X_j) - \lambda_0 G(X_0), X_j \in \mathcal{K}, \lambda_j \geq 0, \beta \in \mathbb{R} \right\}. \end{aligned}$$

The range of values of $\underline{Q}(Z)$ as \underline{Q} ranges over all coherent extensions of \underline{P} to \mathcal{L} is the closed interval $[\underline{E}(Z), \underline{E}(Z)]$.⁴ The value of $\underline{Q}(Z)$ is uniquely determined just when $\underline{E}(Z) = \underline{E}(Z)$, which can be reduced (using the second expression for \underline{E}) to the following condition.

3.1.7 Uniqueness theorem

Suppose \underline{P} is a coherent lower prevision on \mathcal{K} . Then \underline{P} has a unique coherent extension to the gamble Z if and only if, for every $\delta > 0$, there are $X_j \in \mathcal{K}$, $Y_j \in \mathcal{K}$, $\lambda_j \geq 0$, $\rho_j \geq 0$ and $\mu \in \mathbb{R}$ such that

$$\sum_{j=1}^m \rho_j G(Y_j) - \delta \leq Z - \mu \leq \lambda_0 G(X_0) - \sum_{j=1}^n \lambda_j G(X_j) + \delta.$$

(In that case $\underline{E}(Z)$ is the unique coherent extension and we can set $\mu = \underline{E}(Z)$).⁵

This condition simplifies when \mathcal{K} is a convex cone, linear space or field of sets.

3.2 EXTENSION FROM A FIELD

3.1.8 Corollary

Suppose \underline{P} is a coherent lower prevision on \mathcal{K} , where \mathcal{K} is a linear space containing constant gambles. Then \underline{E} and \underline{E} are given by $\underline{E}(Z) = \sup \{\underline{P}(Y) : Z \geq Y, Y \in \mathcal{K}\}$ and $\underline{E}(Z) = \inf \{\underline{P}(Y) : Z \leq Y, Y \in \mathcal{K}\}$, for all $Z \in \mathcal{L}$. There is a unique coherent extension of \underline{P} to Z if and only if, for every $\delta > 0$, there are $X \in \mathcal{K}$ and $Y \in \mathcal{K}$ such that $X \leq Z \leq Y$ and $\underline{P}(Y) - \underline{P}(X) \leq \delta$.

Proof. When $X_j \in \mathcal{K}$ and $\lambda_j \geq 0$, we have

$$\sum_{j=1}^n \lambda_j G(X_j) - \lambda_0 G(X_0) \geq -G\left(\lambda_0 X_0 - \sum_{j=1}^n \lambda_j X_j\right) \quad \text{by super-linearity of } \underline{P}.$$

Hence $\underline{E}(Z) = \inf \{\beta : \beta - Z \geq -G(X), X \in \mathcal{K}\} = \inf \{\underline{P}(Y) : Z \leq Y, Y \in \mathcal{K}\}$ by taking $Y = G(X) + \beta$. The extension is unique just when $\underline{E}(Z) \geq \underline{E}(Z)$, which reduces to the stated condition using the formula for $\underline{E}(Z)$ in Theorem 3.1.4.

3.1.9 Corollary

Suppose \underline{P} is a coherent lower probability on a field of events \mathcal{A} . Then $\underline{E}(B) = \sup \{\underline{P}(A) : B \supset A, A \in \mathcal{A}\}$ and $\underline{E}(B) = \inf \{\underline{P}(A) : A \supset B, A \in \mathcal{A}\}$ for all events B . There is a unique coherent extension of \underline{P} to an event B if and only if, for every $\delta > 0$, there are events A and C in \mathcal{A} such that $A \supset B \supset C$ and $\underline{P}(A) - \underline{P}(C) \leq \delta$. If \mathcal{A} is a σ -field, this holds if and only if there are A and C in \mathcal{A} such that $A \supset B \supset C$ and $\underline{P}(A) = \underline{P}(C)$.

Proof. Suppose $A_j \in \mathcal{A}$, $\lambda_j \geq 0$ and $\beta - B \geq \sum_{j=1}^n \lambda_j G(A_j) - \lambda_0 G(A_0)$. Let $Y = \beta + \lambda_0 G(A_0) - \sum_{j=1}^n \lambda_j G(A_j)$ and $A = \{\omega : Y(\omega) \geq 1\}$. As in the proof of Theorem 3.1.5, $A \in \mathcal{A}$ and $B \leq A \leq Y$ so $A \supset B$. By coherence of \underline{P} ,

$$\begin{aligned} 0 &\leq \sup \left[G(A) + \sum_{j=1}^n \lambda_j G(A_j) - \lambda_0 G(A_0) \right] = \sup [A - \underline{P}(A) + \beta - Y] \\ &\leq \beta - \underline{P}(A), \text{ so } \underline{P}(A) \leq \beta. \end{aligned}$$

It follows that

$$\begin{aligned} \underline{E}(B) &= \inf \left\{ \beta : \beta - B \geq \sum_{j=1}^n \lambda_j G(A_j) - \lambda_0 G(A_0), A_j \in \mathcal{A}, \lambda_j \geq 0 \right\} \\ &\geq \inf \{\underline{P}(A) : A \supset B, A \in \mathcal{A}\}. \end{aligned}$$

The reverse inequality holds, by taking $n = 0$, $A_0 = A$, $\beta = \underline{P}(A)$. The condition for unique extension is equivalent to $\underline{E}(B) \geq \underline{E}(B)$.

If \mathcal{A} is a σ -field and A_n, C_n are sequences in \mathcal{A} such that $A_n \supset B \supset C_n$ and $\underline{P}(A_n) - \underline{P}(C_n) \leq n^{-1}$, let $A = \bigcap_{n=1}^{\infty} A_n$ and $C = \bigcup_{n=1}^{\infty} C_n$. ◆

3.2 Extension from a field

This section is concerned with extending a coherent lower probability, defined on a field of events \mathcal{A} , to a coherent lower prevision on the class of \mathcal{A} -measurable gambles. This is an especially important type of extension because it will often be convenient to specify beliefs about Ω through probabilities of events and then investigate their implications for other gambles.

When the specified probabilities are additive, their natural extension to the class of measurable gambles is a linear prevision and agrees with the usual definition of linear expectation (section 3.2.1). When the probabilities are countably additive, on a σ -field the extension coincides with the classical construction of an integral from a measure.

In the general case of a coherent lower probability P defined on a field \mathcal{A} , there are three ways to compute the natural extension. One is to use the basic definition 3.1.1 or equivalent formulas 3.1.3, but that will usually be difficult. A second method is to find the additive probabilities which dominate P on \mathcal{A} , construct their expectations for \mathcal{A} -measurable gambles, and obtain the natural extension as the lower envelope of these expectations (Theorem 3.4.1). When P has a regularity property of 2-monotonicity, it is often easier to construct the natural extension by integrating upper and lower distribution functions. This construction is described in section 3.2.4.

3.2.1 Linear expectation

First we show that linear expectation, a basic concept in classical probability theory, is a special type of natural extension. It is just the natural extension of an additive probability from a field \mathcal{A} to the class of \mathcal{A} -measurable gambles.

Suppose that an additive probability P is defined on a field of events \mathcal{A} . The natural extension of P to all gambles is defined by

$$\begin{aligned} E(X) &= \sup \left\{ \alpha : X - \alpha \geq \sum_{j=1}^n \lambda_j G(A_j) \quad \text{for } n \geq 0, A_j \in \mathcal{A}, \lambda_j \geq 0 \right\}, \\ &= \sup \left\{ \sum_{j=1}^n \lambda_j P(A_j) + \mu : X \geq \sum_{j=1}^n \lambda_j A_j + \mu, n \geq 0, A_j \in \mathcal{A}, \lambda_j \geq 0, \mu \in \mathbb{R} \right\} \end{aligned}$$

using Lemma 3.1.3(b). Here $G(A) = A - P(A) = -G(A^\complement)$ by additivity of P , hence the constants λ_j can be allowed to take any real values. We can also replace μ in the second formula by $\mu P(\Omega)$ and $\mu \Omega$, since $\Omega \in \mathcal{A}$ and $P(\Omega) = 1$. This gives the expression

$$\underline{E}(X) = \sup \left\{ \sum_{j=1}^n \lambda_j P(A_j) : X \geq \sum_{j=1}^n \lambda_j A_j, n \geq 0, A_j \in \mathcal{A}, \lambda_j \in \mathbb{R} \right\},$$

3.2 EXTENSION FROM A FIELD

which (for \mathcal{A} -measurable gambles X) is often taken as the definition of the **expectation** of X under P . (This is usually regarded as an integral with respect to the measure P , and denoted by $\int X dP$ or $\int X(\omega) P(d\omega)$ rather than $E(X)$.) A gamble X is said to be \mathcal{A} -measurable when $\{\omega : X(\omega) > \mu\}$ and $\{\omega : X(\omega) < \mu\}$ are in \mathcal{A} for every real μ . Let $\mathcal{K}(\mathcal{A})$ denote the set of all \mathcal{A} -measurable gambles. (This is a linear space whenever \mathcal{A} is a σ -field, but not necessarily when \mathcal{A} is a field.) The following theorem establishes that the natural extension \underline{E} is a linear prevision on $\mathcal{K}(\mathcal{A})$.

3.2.2 Expectation theorem

Let P be an additive probability on a field \mathcal{A} , and \underline{E} its natural extension to \mathcal{L} . Then \underline{E} is a linear prevision on $\mathcal{K}(\mathcal{A})$. Also \underline{E} is the unique coherent lower prevision (hence, the unique linear prevision) on $\mathcal{K}(\mathcal{A})$ that agrees with P on \mathcal{A} .

Proof. Let $\mathcal{T} = \{X \in \mathcal{L} : \underline{E}(X) = \bar{E}(X)\}$ be the class of gambles for which \underline{E} agrees with its conjugate upper prevision \bar{E} . Since \underline{E} is a coherent lower prevision, it follows from the basic properties that \mathcal{T} is a linear space. Since P is additive and \underline{E} agrees with P on \mathcal{A} (by 3.1.2(d)), \mathcal{T} contains \mathcal{A} . So \mathcal{T} contains all \mathcal{A} -simple gambles, of the form $X = \sum_{j=1}^m \lambda_j A_j$ where $A_j \in \mathcal{A}$ and $\lambda_j \in \mathbb{R}$. Also \mathcal{T} is closed under supremum norm, since if $X_n \in \mathcal{T}$ with $\sup |X_n - X| \rightarrow 0$ then $\underline{E}(X_n) \rightarrow \underline{E}(X)$ and $\bar{E}(X_n) \rightarrow \bar{E}(X)$ by property 2.6.1(l), so $\underline{E}(X) = \bar{E}(X)$. But every $X \in \mathcal{K}(\mathcal{A})$ can be written as a uniform limit of \mathcal{A} -simple gambles X_n , by defining $X_n = \sum_{j=1}^m \lambda_j A_j$ where $A_j = \{\omega : \lambda_{j-1} < X(\omega) \leq \lambda_j\}$, $\lambda_j - \lambda_{j-1} = n^{-1}$, $\lambda_0 < \inf X$ and $\lambda_m \geq \sup X$, so that $\sup |X_n - X| \leq n^{-1}$. This shows that \mathcal{T} contains $\mathcal{K}(\mathcal{A})$. Since \underline{E} is self-conjugate on \mathcal{T} , it is a linear prevision on \mathcal{T} by Theorem 2.8.2. The uniqueness of the extension to $\mathcal{K}(\mathcal{A})$ follows from 3.1.2(f): if \underline{Q} is any coherent extension of P to $\mathcal{K}(\mathcal{A})$ then $\underline{E} \leq \underline{Q} \leq \bar{Q} \leq \bar{E} = \underline{E}$, so $\underline{Q} = \underline{E}$. ◆

3.2.3 Corollary

Suppose that P is an additive probability defined on all subsets of Ω . Then its natural extension \underline{E} is a linear prevision on \mathcal{L} , and is the unique coherent extension of P to \mathcal{L} (as either a linear prevision or a lower prevision).

Proof. Let \mathcal{A} contain all subsets of Ω . All gambles are \mathcal{A} -measurable, so $\mathcal{K}(\mathcal{A}) = \mathcal{L}$ in the previous theorem. ◆

3.2.4 Upper and lower distribution functions

When P is an additive probability on a field \mathcal{A} , its natural extension to $\mathcal{K}(\mathcal{A})$ can be written in the alternative form $\underline{E}(X) = \int_{-\infty}^{\infty} x dF_X(x)$, where F_X

is the **distribution function** of X under P , defined by $F_X(x) = P(\{\omega: X(\omega) \leq x\})$. To see that, let X be an \mathcal{A} -simple gamble and write $X = \sum_{j=1}^n \lambda_j A_j$ where $A_j \in \mathcal{A}$, $\{A_1, \dots, A_n\}$ is a partition of Ω , and $\lambda_1 < \lambda_2 < \dots < \lambda_n$. By linearity of E on $\mathcal{K}(\mathcal{A})$, $E(X) = \sum_{j=1}^n \lambda_j E(A_j) = \sum_{j=1}^n \lambda_j P(A_j) = \int_{-\infty}^{\infty} x dF_X(x)$. Since each $X \in \mathcal{K}(\mathcal{A})$ can be uniformly approximated from above and below by \mathcal{A} -simple gambles, the formula extends to all $X \in \mathcal{K}(\mathcal{A})$.

This formulation generalizes to an important class of lower probabilities, those which are 2-monotone. A coherent lower probability \underline{P} , defined on a field \mathcal{A} , is said to be **2-monotone** if it satisfies

$$\underline{P}(A \cup B) + \underline{P}(A \cap B) \geq \underline{P}(A) + \underline{P}(B) \quad \text{whenever } A \in \mathcal{A} \text{ and } B \in \mathcal{A}.$$

(There is equality here when \underline{P} is an additive probability, so all additive probabilities are 2-monotone.)¹

Define the **upper and lower distribution functions** of X under \underline{P} by

$$\bar{F}_X(x) = \bar{P}(\{\omega: X(\omega) \leq x\}) = 1 - \underline{P}(\{\omega: X(\omega) > x\})$$

and $F_X(x) = \underline{P}(\{\omega: X(\omega) \leq x\})$.

When \underline{P} is 2-monotone on \mathcal{A} , it can be shown² that its natural extension to $\mathcal{K}(\mathcal{A})$ is given by the formulas

$$\begin{aligned} \underline{E}(X) &= \int_{-\infty}^{\infty} x d\bar{F}_X(x) = \int_0^{\infty} 1 - \bar{F}_X(x) dx - \int_{-\infty}^0 \bar{F}_X(x) dx \\ &= \sup X - \int_{\inf X}^{\sup X} \bar{F}_X(x) dx = \inf X + \int_{\inf X}^{\sup X} 1 - \bar{F}_X(x) dx. \end{aligned}$$

The conjugate upper prevision $\bar{E}(X) = \int_{-\infty}^{\infty} x d\underline{F}_X(x)$ satisfies similar formulas with $\bar{F}_X(x)$ replaced by $\underline{F}_X(x)$. Because upper and lower distribution functions can be easily computed from \underline{P} , these formulas can greatly simplify computation of the natural extension, as in the following examples.

3.2.5 Pari-mutuel model

Consider the *pari-mutuel* (P_0, δ) upper and lower probabilities, defined in section 2.9.3 by $\bar{P}(A) = \min\{(1+\delta)P_0(A), 1\}$ and $\underline{P}(A) = \max\{(1+\delta)P_0(A) - \delta, 0\}$. Because P_0 is additive, \underline{P} is 2-monotone. Writing F_X for the distribution function of X under P_0 , the upper and lower distribution functions are $\bar{F}_X(x) = \min\{(1+\delta)F_X(x), 1\}$ and $\underline{F}_X(x) = \max\{(1+\delta)F_X(x) - \delta, 0\}$.

Write $\alpha = \inf X$, $\beta = \sup X$, $\tau = \delta/(1+\delta)$, and let x_τ denote the (upper) τ -quantile of X under P_0 , defined to be the supremum value of x such that $F_X(x) \leq \tau$. Let \underline{E} be the natural extension of \underline{P} to \mathcal{L} , and \bar{E} its conjugate

3.2 EXTENSION FROM A FIELD

upper prevision. Then \bar{E} is given by

$$\begin{aligned} \bar{E}(X) &= \alpha + \int_{\alpha}^{\beta} 1 - \bar{F}_X(x) dx = \alpha + \int_{\alpha}^{\beta} \min\{(1+\delta)(1 - F_X(x)), 1\} dx \\ &= \alpha + \int_{\alpha}^{x_\tau} dx + \int_{x_\tau}^{\beta} (1+\delta)(1 - F_X(x)) dx \\ &= x_\tau + (1+\delta) \int_{x_\tau}^{\beta} 1 - F_X(x) dx = x_\tau + (1+\delta)P_0((X - x_\tau)^+). \end{aligned}$$

Similarly,

$$\underline{E}(X) = x_{1-\tau} - (1+\delta) \int_{\alpha}^{x_{1-\tau}} F_X(x) dx = x_{1-\tau} - (1+\delta)P_0((x_{1-\tau} - X)^+).$$

If X has a continuous distribution function F_X then $P_0(\{\omega: X(\omega) > x_\tau\}) = 1 - \tau = (1+\delta)^{-1}$, from which it follows that $\bar{E}(X) = P_0(X|X > x_\tau)$, the prevision of X conditional on the event that it exceeds its τ -quantile.³ Similarly $\underline{E}(X) = P_0(X|X < x_{1-\tau})$.

3.2.6 Zero-one valued probabilities

As in section 2.9.8, let \underline{P} be the 0–1 valued lower probability generated by the filter \mathcal{A} , defined by $\underline{P}(A) = 1$ if $A \in \mathcal{A}$, $\underline{P}(A) = 0$ otherwise. We wish to find the natural extension of \underline{P} to $\mathcal{L}(\Omega)$. Note that \underline{P} is a 2-monotone lower probability, since if $\underline{P}(A) > 0$ and $\underline{P}(B) > 0$ then $A \in \mathcal{A}$ and $B \in \mathcal{A}$, so $A \cap B \in \mathcal{A}$ and $\underline{P}(A \cap B) = 1$. The natural extension of \underline{P} can therefore be computed from the upper and lower distribution functions, which are also 0–1 valued.

Given $X \in \mathcal{L}$, let \underline{x} be the supremum value of x for which $\{\omega: X(\omega) > x\} \in \mathcal{A}$, and let \bar{x} be the infimum value of x for which $\{\omega: X(\omega) \leq x\} \in \mathcal{A}$. Clearly $\underline{x} \leq \bar{x}$, and $\underline{x} = \bar{x}$ when \mathcal{A} is an ultrafilter. The upper and lower distribution functions of X satisfy $\bar{F}_X(x) = 0$ if $x < \underline{x}$, $\bar{F}_X(x) = 1$ if $x > \bar{x}$, $F_X(x) = 0$ if $x < \bar{x}$, and $\underline{F}_X(x) = 1$ if $x > \underline{x}$. Hence the natural extension to \mathcal{L} is $\underline{E}(X) = \int_{-\infty}^{\infty} x d\bar{F}_X(x) = \underline{x}$ with $\bar{E}(X) = \int_{-\infty}^{\infty} x d\underline{F}_X(x) = \bar{x}$.⁴

When \mathcal{A} is an ultrafilter, \underline{P} is an additive probability and its expectation $\underline{E} = \bar{E}$ is a linear prevision on \mathcal{L} . At the other extreme, when $\mathcal{A} = \{\Omega\}$, \underline{P} is the vacuous lower probability and its natural extension \underline{E} is the vacuous lower prevision (since then $\underline{x} = \inf X$ and $\bar{x} = \sup X$).

3.2.7 Inner and outer Lebesgue measure

In section 2.9.6 we defined inner and outer Lebesgue measure for all subsets A of $[0, 1]$ by $\underline{P}(A) = \sup\{v(B): A \supseteq B, B \in \mathbb{B}\}$ and $\bar{P}(A) = \inf\{v(B): B \supseteq A, B \in \mathbb{B}\}$, where v is Lebesgue measure on the Borel σ -field \mathbb{B} . We can apply the

preceding results on natural extension to deduce the following facts about inner and outer Lebesgue measure.

1. \underline{P} is the natural extension of v to all events, and \bar{P} is its conjugate upper probability (see Theorem 3.1.5).
2. The supremum and infimum in the definitions are attained by some $B \in \mathcal{B}$. (See the remarks following Theorem 3.1.5.)
3. \underline{P} is the minimal coherent lower probability defined on all subsets of Ω that agrees with v on \mathcal{B} (Properties 3.1.2).
4. As \underline{Q} ranges over all coherent lower probabilities which extend v , $\underline{Q}(A)$ takes all values in the closed interval $[\underline{P}(A), \bar{P}(A)]$ (Corollary 3.1.9). (In fact, all such values can be attained by additive probabilities which extend v : see Corollary 3.4.3.)
5. The Lebesgue-measurable sets, for which $\underline{P}(A) = \bar{P}(A)$, are just the sets A such that there are $B \in \mathcal{B}$ and $C \in \mathcal{B}$ with $B \supset A \supset C$ and $v(B) = v(C)$ (Corollary 3.1.9). By countable additivity of v on \mathcal{B} , the class of Lebesgue-measurable sets is a σ -field which contains all sets with outer measure zero. (It is called the **completion** of \mathcal{B} .)
6. There is a unique linear prevision P_0 , defined on the linear space \mathcal{K} of \mathcal{B} -measurable gambles, which agrees with v on \mathcal{B} . (P_0 is just the natural extension of v to \mathcal{K}) (Theorem 3.2.2).
7. The natural extension of v to all gambles is the Lebesgue lower prevision $\underline{E}(X) = \sup\{P_0(Y): Y \leq X, Y \in \mathcal{K}\}$ (Theorem 3.1.4). On the class of Lebesgue-measurable gambles, \underline{E} is a linear prevision and is just the Lebesgue integral.
8. Inner Lebesgue measure is 2-monotone, so the natural extension \underline{E} is also given by $\underline{E}(X) = \int_{-\infty}^{\infty} x d\bar{F}_x(x)$, where \bar{F}_x is the upper distribution function generated by \underline{P} (section 3.2.4).

3.3 Lower envelopes of linear previsions

Next we look more closely at the relationship between a lower prevision and its dominating linear previsions. Theorem 2.6.3 shows that for a lower prevision \underline{P} on \mathcal{K} to avoid sure loss it is sufficient that there is a linear prevision P which **dominates** \underline{P} on \mathcal{K} , i.e. $P(X) \geq \underline{P}(X)$ for all $X \in \mathcal{K}$. Moreover, for \underline{P} to be coherent it is sufficient that \underline{P} is a **lower envelope** of linear previsions, i.e. there is a class \mathcal{M} of linear previsions such that $\underline{P}(X) = \inf\{P(X): P \in \mathcal{M}\}$ for all $X \in \mathcal{K}$. In this section we show that these conditions are necessary as well as sufficient for avoiding sure loss and coherence.

3.3.1 Notation

Let \mathcal{P} denote the class of all linear previsions on domain $\mathcal{L} = \mathcal{L}(\Omega)$. For any lower prevision \underline{P} on arbitrary domain \mathcal{K} , $\mathcal{M}(\underline{P})$ denotes the class of

all linear previsions that dominate \underline{P} on \mathcal{K} ,

$$\mathcal{M}(\underline{P}) = \{P \in \mathcal{P}: P(X) \geq \underline{P}(X) \text{ for all } X \in \mathcal{K}\}.$$

If \underline{P} and \underline{Q} are lower previsions on the same domain and $\underline{P} \geq \underline{Q}$ then $\mathcal{M}(\underline{Q}) \supset \mathcal{M}(\underline{P})$.

If \underline{P} is the lower envelope of a class of linear previsions \mathcal{M} then every P in \mathcal{M} dominates \underline{P} , so $\mathcal{M}(\underline{P}) \supset \mathcal{M}$. It follows that \underline{P} is the lower envelope of some class of linear previsions if and only if it is the lower envelope of $\mathcal{M}(\underline{P})$. (In that case we will simply call \underline{P} a lower envelope.) In examining whether \underline{P} is a lower envelope we need consider only the class $\mathcal{M}(\underline{P})$.

In general, $\mathcal{M}(\underline{P})$ may be empty. The following lemma in effect shows that $\mathcal{M}(\underline{P})$ is non-empty if and only if \underline{P} avoids sure loss. It is then easy to show that all coherent lower previsions are lower envelopes. Whereas all the preceding results have relied only on elementary properties of linear spaces, this lemma uses the separating hyperplane theorem to guarantee the existence of a positive linear functional that ‘separates’ the zero gamble from a class of desirable gambles.¹

3.3.2 Separation lemma

Let \mathcal{D} be an arbitrary subset of \mathcal{L} . The following three conditions are equivalent:

- (a) $\sup \sum_{j=1}^n X_j \geq 0$ whenever $n \geq 1$ and $X_1, \dots, X_n \in \mathcal{D}$
- (b) there is a linear prevision P on domain \mathcal{L} such that $P(X) \geq 0$ for all $X \in \mathcal{D}$
- (c) there is a linear prevision P on domain \mathcal{D} such that $P(X) \geq 0$ for all $X \in \mathcal{D}$.

*Proof.*² Suppose (a) holds. Let $\mathcal{V} = \{Y \in \mathcal{L}: Y \geq \sum_{j=1}^n \lambda_j X_j \text{ for some } n \geq 0, X_j \in \mathcal{D}, \lambda_j \geq 0\}$. Clearly \mathcal{V} is a convex subset of \mathcal{L} . (In fact, \mathcal{V} is the smallest convex cone that contains \mathcal{D} and all non-negative gambles.) To apply the separating hyperplane theorem, consider the topology on \mathcal{L} generated by the supremum norm $\|X\| = \sup |X|$. Under this topology, \mathcal{L} is a linear topological space (Appendix D) and \mathcal{V} has non-empty topological interior $\text{int}(\mathcal{V})$, e.g. $1 \in \text{int}(\mathcal{V})$. When $Y \in \mathcal{V}$, $\sup Y \geq \sup \sum_{j=1}^n \lambda_j X_j \geq 0$, using (a) and the argument in Lemma 2.4.4(a). If $Y \in \text{int}(\mathcal{V})$ then there is $\delta > 0$ such that $Y - \delta \in \mathcal{V}$, and then $\sup Y \geq \delta > 0$. This proves that the zero gamble is not in $\text{int}(\mathcal{V})$. By a version of the separating hyperplane theorem (Appendix E1), there is a non-zero linear functional P on \mathcal{L} such that $P(Y) \geq 0$ for all $Y \in \mathcal{V}$. (This ‘separates’ the zero gamble from \mathcal{V} .)

If $Y \geq 0$ then $Y \in \mathcal{V}$ and $P(Y) \geq 0$. Thus P is a positive linear functional on \mathcal{L} . Since P is non-zero, $P(1)$ is positive so we can renormalize P to have unit norm, $P(1) = 1$. Then, by Corollary 2.8.5, P is a linear prevision on \mathcal{L} , and (b) holds because $\mathcal{V} \supset \mathcal{D}$. Thus (a) implies (b).

Clearly (b) implies (c), since the restriction of a linear prevision to a smaller domain \mathcal{D} is still a linear prevision. To see that (c) implies (a), let $X_1, \dots, X_n \in \mathcal{D}$. Then $\sup \sum_{j=1}^n X_j \geq \sum_{j=1}^n P(X_j)$ by the definition of linear prevision (2.8.1), and this is non-negative by (c). \blacklozenge

The lemma can now be applied to characterize the lower previsions \underline{P} on domain \mathcal{K} that avoid sure loss, by taking $\mathcal{D} = \{G(X): X \in \mathcal{K}\}$ to be the class of marginally desirable gambles, and those \underline{P} that are coherent, by considering $\mathcal{D}(X_0) = \mathcal{D} \cup \{-G(X_0)\}$ for each $X_0 \in \mathcal{K}$.

3.3.3 Lower envelope theorem³

Suppose \underline{P} is a lower prevision on domain \mathcal{K} , where \mathcal{K} is an arbitrary subset of \mathcal{L} .

- (a) \underline{P} avoids sure loss if and only if $\mathcal{M}(\underline{P})$ is non-empty (i.e., if and only if \underline{P} is dominated by some linear prevision).
- (b) \underline{P} is coherent if and only if it is the lower envelope of $\mathcal{M}(\underline{P})$ (i.e., if and only if \underline{P} is the lower envelope of some class of linear previsions).

Proof. (a) Apply Lemma 3.3.2 to $\mathcal{D} = \{G(X): X \in \mathcal{K}\}$. By Definition 2.4.1, \underline{P} avoids sure loss if and only if condition 3.3.2(a) holds. That is equivalent to the existence of $P \in \mathcal{P}$ such that $P(G(X)) \geq 0$ for all $X \in \mathcal{K}$. But $P(G(X)) = P(X - \underline{P}(X)) = P(X) - \underline{P}(X)$, so $P \in \mathcal{M}(\underline{P})$.

(b) For $X_0 \in \mathcal{K}$, apply 3.3.2 to $\mathcal{D}(X_0) = \{G(X): X \in \mathcal{K}\} \cup \{-G(X_0)\}$. The coherence condition 2.5.1 holds for fixed X_0 if and only if 3.3.2(a) holds for $\mathcal{D}(X_0)$. That is equivalent to the existence of $P \in \mathcal{P}$ such that $P(G(X)) \geq 0$ for all $X \in \mathcal{K}$ and $P(G(X_0)) \leq 0$, i.e. $P \in \mathcal{M}(\underline{P})$ and $P(X_0) = \underline{P}(X_0)$. Thus \underline{P} is coherent if and only if, for every $X_0 \in \mathcal{K}$, there is $P \in \mathcal{M}(\underline{P})$ with $P(X_0) = \underline{P}(X_0)$, i.e. $\underline{P}(X_0) = \min\{P(X_0): P \in \mathcal{M}(\underline{P})\}$. \blacklozenge

The proof makes clear the relationship between coherence and dominating linear previsions. The lower prevision \underline{P} incurs sure loss just when it has no dominating linear previsions, and \underline{P} fails the coherence condition 2.5.1 for a particular X_0 just when $\underline{P}(X_0)$ cannot be attained by any dominating linear prevision. Note that, for all X in the domain of a coherent lower prevision \underline{P} , $\underline{P}(X) = \min\{P(X): P \in \mathcal{M}(\underline{P})\}$. (The infimum is actually attained by some dominating linear prevision.)

We have required the linear previsions in $\mathcal{M}(\underline{P})$ to be defined on \mathcal{L} , but it is easy to see that Theorem 3.3.3 remains valid if they are defined only on \mathcal{K} . If P is a linear prevision on \mathcal{L} that dominates \underline{P} on \mathcal{K} then the restriction of P to \mathcal{K} is also a linear prevision that dominates \underline{P} on \mathcal{K} . So if \underline{P} avoids sure loss then there is a linear prevision on \mathcal{K} that dominates \underline{P} on \mathcal{K} . The converse holds by Corollary 2.8.6. Similarly, \underline{P} is coherent if

3.3 LOWER ENVELOPES OF LINEAR PREVISIONS

and only if it is the lower envelope of a class of linear previsions each with domain \mathcal{K} .⁴ For instance, the additive probabilities in the following corollary can be regarded as defined on the domain \mathcal{A} of \underline{P} , or on all events, or as linear previsions defined on all gambles. This result, which is a direct application of Theorem 3.3.3, characterizes avoiding sure loss and coherence for lower probabilities.

3.3.4 Corollary

Suppose \underline{P} is a lower prevision on domain \mathcal{K} , where \mathcal{K} is an arbitrary class of subsets of Ω .

- (a) \underline{P} avoids sure loss if and only if there is an additive probability that dominates \underline{P} on \mathcal{K} .
- (b) \underline{P} is coherent if and only if it is the lower envelope of a class of additive probabilities.

3.3.5 Examples

Here we identify the class $\mathcal{M}(\underline{P})$ for each of the coherent lower previsions \underline{P} defined in section 2.9.

(a) Vacuous previsions (2.9.1)

$\mathcal{M}(\underline{P}) = \mathcal{P}$, the class of all linear previsions on \mathcal{L} .

(b) Linear–vacuous mixtures (2.9.2)

$\mathcal{M}(\underline{P}) = \{(1 - \delta)P_0 + \delta P: P \in \mathcal{P}\}$, the class of all convex combinations of P_0 with an arbitrary linear prevision P , with weights $1 - \delta$ and δ .

(c) Pari-mutuel models (2.9.3)

$\mathcal{M}(\underline{P}) = \{P \in \mathcal{P}: P(A) \leq (1 + \delta)P_0(A) \text{ for all events } A\}$, the class of all linear previsions under which the probabilities of events do not exceed a constant multiple of P_0 .

(d) Constant odds-ratio models (2.9.4)

$\mathcal{M}(\underline{P}) = \{P \in \mathcal{P}: P(A)/P(B) \geq (1 - \tau)P_0(A)/P_0(B) \text{ for all events } A, B\}$, the class of linear previsions P for which the ratios of probabilities $P(A)/P(B)$ are all within a constant multiple of the corresponding ratios under P_0 .

(e) Uniform distributions on the positive integers (2.9.5)

$\mathcal{M}(\underline{Q})$ is the class of all shift-invariant linear previsions on $\mathcal{L}(\mathbb{Z}^+)$. (See section 3.5.7.)

(f) Inner and outer Lebesgue measure (2.9.6)

$\mathcal{M}(\underline{P})$ is the class of expectations for finitely additive extensions of Lebesgue measure to all subsets of $[0, 1]$. (See section 3.4.2.)

(g) Uniform distributions on the real line (2.9.7)

$\mathcal{M}(Q)$ is the class of all translation-invariant linear previsions which can be obtained as generalized limits of uniform distributions on intervals whose length tends to infinity. When the intervals are restricted to be centred at zero, we obtain the smaller class $\mathcal{M}(\underline{P})$.

(h) Zero-one valued probabilities (2.9.8)

$\mathcal{M}(\underline{P}) = \{P \in \mathcal{P}: P(A) = 1 \text{ for all } A \in \mathcal{A}\}$, the class of all linear previsions assigning probability one to all sets in the filter \mathcal{A} .

3.4 Linear extension

The lower envelope theorem 3.3.3 can be used to give a simple characterization of the natural extension in terms of dominating linear previsions.

3.4.1 Natural extension theorem¹

Suppose that the lower prevision $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss. Then its natural extension \underline{E} is the lower envelope of $\mathcal{M}(\underline{P})$, i.e. $\underline{E}(X) = \min\{\underline{P}(X): P \in \mathcal{M}(\underline{P})\}$ for all $X \in \mathcal{L}$, and $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{P})$.

Proof. Since \underline{P} avoids sure loss, $\mathcal{M}(\underline{P})$ is non-empty by Theorem 3.3.3(a). Define $\underline{Q}(X) = \inf\{\underline{P}(X): P \in \mathcal{M}(\underline{P})\}$ for all $X \in \mathcal{L}$, so \underline{Q} is the lower envelope of $\mathcal{M}(\underline{P})$. Then $\mathcal{M}(\underline{Q}) \supset \mathcal{M}(\underline{P})$, since $\mathcal{M}(\underline{Q})$ is the largest class of which \underline{Q} is the lower envelope. Also \underline{Q} is coherent (as a lower envelope) and $\underline{Q} \geq \underline{P}$ on \mathcal{K} . By property 3.1.2(e) of natural extension, $\underline{Q} \geq \underline{E}$ on \mathcal{L} and $\mathcal{M}(\underline{E}) \supset \mathcal{M}(\underline{Q})$. But $\underline{E} \geq \underline{P}$ on \mathcal{K} by 3.1.2(c), hence $\mathcal{M}(\underline{P}) \supset \mathcal{M}(\underline{E})$. Thus $\mathcal{M}(\underline{P}) \supset \mathcal{M}(\underline{E}) \supset \mathcal{M}(\underline{Q}) \supset \mathcal{M}(\underline{P})$, giving $\mathcal{M}(\underline{P}) = \mathcal{M}(\underline{E}) = \mathcal{M}(\underline{Q})$. But \underline{E} , coherent by 3.1.2(b), is the lower envelope of $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{P})$ by 3.3.3(b). ◆

The theorem suggests a useful method for computing the natural extension of a lower prevision \underline{P} . Find the linear previsions which dominate \underline{P} on its domain \mathcal{K} , form all their extensions to linear previsions on \mathcal{L} , and construct \underline{E} as the lower envelope of the class of all such extensions $\mathcal{M}(\underline{P})$. (Theorem 3.6.2 shows that only the extreme points of $\mathcal{M}(\underline{P})$ are needed.)

In order to construct \underline{E} in this way, we need to know how to extend a linear prevision from \mathcal{K} to \mathcal{L} . Several characterizations of the linear extensions will be given in this section. The next theorem shows that a linear prevision P defined on an arbitrary domain \mathcal{K} can always be extended to

3.4 LINEAR EXTENSION

a linear prevision defined on all gambles.² When $\mathcal{K} = -\mathcal{K}$, the linear extensions are just the linear previsions in $\mathcal{M}(P)$.

3.4.2 Linear extension theorem

Suppose P is a linear prevision on \mathcal{K} . Then P has linear extensions to \mathcal{L} . If $\mathcal{K} = -\mathcal{K}$ then the class of all linear extensions to \mathcal{L} is $\mathcal{M}(P)$. Generally the class of all linear extensions is $\mathcal{M}(P')$, where P' is the linear extension of P to $\mathcal{K} \cup (-\mathcal{K})$, defined for $X \in -\mathcal{K}$ by $P'(X) = -P(-X)$.³

Proof. Assume $\mathcal{K} = -\mathcal{K}$. If Q is a linear extension of P to \mathcal{L} then $Q = P$ on \mathcal{K} , so $Q \in \mathcal{M}(P)$. Conversely, if $Q \in \mathcal{M}(P)$ then, for all $X \in \mathcal{K}$, $Q(X) \geq P(X) = -P(-X) \geq -Q(-X) = Q(X)$ using Theorem 2.8.2, so $Q(X) = P(X)$ and Q is a linear extension of P . Thus the class of all linear extensions is $\mathcal{M}(P)$, which is non-empty by 3.3.3(a) since P avoids sure loss.

For the general result, use Definition 2.8.1 and Theorem 2.8.2 to show that P' is a linear prevision on $\mathcal{K}' = \mathcal{K} \cup (-\mathcal{K})$, which satisfies $\mathcal{K}' = -\mathcal{K}'$. The class of linear extensions of P' to \mathcal{L} is $\mathcal{M}(P')$ by the first result. The linear extensions of P are just the linear extensions of P' . ◆

For example, an additive probability defined on an arbitrary collection of subsets of Ω can always be extended to an additive probability on all subsets. In the case where P is Lebesgue measure defined on the Borel- or Lebesgue-measurable subsets of $[0, 1]$, there are (finitely) additive extensions of P to all subsets of $[0, 1]$. The lower envelope of all such extensions is the natural extension of P , which is inner Lebesgue measure \underline{P} .⁴ For every subset A , the range of values $Q(A)$ taken by additive extensions is $[\underline{P}(A), \bar{P}(A)]$, where \bar{P} is outer Lebesgue measure.

In general, the range of values taken by linear extensions is determined by the natural extension, as follows.

3.4.3 Corollary

Suppose P is a linear prevision on \mathcal{K} . If $\mathcal{K} = -\mathcal{K}$, let \underline{E} be the natural extension of P to \mathcal{L} ; in general let \underline{E} be the natural extension of P' defined in 3.4.2. Then the class of all linear extensions of P to \mathcal{L} is $\mathcal{M}(\underline{E})$. For every gamble X , as Q ranges over all linear extensions of P , $Q(X)$ takes all values in the closed interval $[\underline{E}(X), \bar{E}(X)]$.

Proof. We can assume $\mathcal{K} = -\mathcal{K}$. (Otherwise replace P by P' .) The class of all linear extensions is $\mathcal{M}(P) = \mathcal{M}(\underline{E})$ by 3.4.1 and 3.4.2. Given $X \in \mathcal{L}$, $\underline{E}(X) \leq Q(X) \leq \bar{E}(X)$ for all $Q \in \mathcal{M}(\underline{E})$. By 3.3.3 there is some $Q_1 \in \mathcal{M}(\underline{E})$ with $Q_1(X) = \underline{E}(X)$. Similarly, there is $Q_2 \in \mathcal{M}(\underline{E})$ with $Q_2(X) = \bar{E}(X)$. Since $\mathcal{M}(\underline{E})$ is convex, all intermediate values can be attained by convex combinations of Q_1 and Q_2 . ◆

In the general case, the natural extension \underline{E} in the corollary can be written, using 3.1.3(b), as

$$\underline{E}(X) = \sup \left\{ \sum_{j=1}^n \lambda_j P(X_j) + \mu : X \geq \sum_{j=1}^n \lambda_j X_j + \mu, n \geq 0, X_j \in \mathcal{K}, \lambda_j \in \mathbb{R}, \mu \in \mathbb{R} \right\}.$$

There is a unique linear extension of P to the linear space \mathcal{F} generated by \mathcal{K} and the constant gambles, defined by $Q(\sum_{j=1}^n \lambda_j X_j + \mu) = \sum_{j=1}^n \lambda_j P(X_j) + \mu$. Then \underline{E} in the corollary is simply the natural extension of Q from \mathcal{F} to \mathcal{L} , $\underline{E}(X) = \sup \{Q(Y) : X \geq Y, Y \in \mathcal{F}\}$, with conjugate upper prevision $\bar{E}(X) = \inf \{Q(Y) : X \leq Y, Y \in \mathcal{F}\}$.

Corollary 3.4.3 is apparently equivalent to a result asserted by de Finetti (1974, sections 3.10.1 and 3.10.7), who calls it ‘the fundamental theorem of probability’. It is interesting that his ‘fundamental theorem’ explicitly involves upper and lower probabilities \bar{E} and \underline{E} , although de Finetti regards these merely as upper and lower bounds for probabilities which could be precisely evaluated.⁵

Using these results we can characterize the gambles for which the linear extension of P is uniquely determined, i.e. those X for which $\underline{E}(X) = \bar{E}(X)$. As in Theorem 3.2.2, the class of such gambles is a linear space which contains \mathcal{F} and is closed under supremum norm. From the above expressions for \underline{E} and \bar{E} , $\underline{E}(X) = \bar{E}(X)$ if and only if, for every $\delta > 0$, there are $Y \in \mathcal{F}$ and $Z \in \mathcal{F}$ such that $Y \geq X \geq Z$ and $Q(Y) - Q(Z) \leq \delta$.⁶

3.4.4 Hahn–Banach extensions

Now consider linear extensions which dominate a given lower prevision. Suppose that \underline{P} is a lower prevision on \mathcal{K} , P is a linear prevision on \mathcal{T} where $\mathcal{L} \supset \mathcal{K} \supset \mathcal{T}$, and P dominates \underline{P} on \mathcal{T} . Under what conditions can P be extended to a linear prevision that dominates \underline{P} on \mathcal{K} ? Since P has a unique linear extension to the linear space generated by \mathcal{T} , we simplify the problem by assuming that \mathcal{T} is a linear space. Extension of P is then possible provided \underline{P} is coherent on \mathcal{K} . In fact, the following weaker conditions are sufficient.

3.4.5 Theorem

Suppose $\mathcal{L} \supset \mathcal{K} \supset \mathcal{T}$, \mathcal{T} is a linear space, \underline{P} is a lower prevision on \mathcal{K} , P is a linear prevision on \mathcal{T} , and $P \geq \underline{P}$ on \mathcal{T} . Then P can be extended to a linear prevision which dominates \underline{P} on \mathcal{K} if and only if

$$\sup \left[\sum_{j=1}^n G(X_j) - X_0 + P(X_0) \right] \geq 0 \text{ whenever } n \geq 0, X_0 \in \mathcal{T} \text{ and } X_1, \dots, X_n \in \mathcal{K}.$$

3.5 INVARIANT LINEAR PREVISIONS

For this to hold it is sufficient that \underline{P} satisfies the coherence condition 2.5.1 whenever $X_0 \in \mathcal{T}$.

Proof. Apply Lemma 3.3.2 with $\mathcal{D} = \{G(X) : X \in \mathcal{K}\} \cup \{Y - P(Y) : Y \in \mathcal{T}\}$. Condition 3.3.2(b) guarantees the existence of a linear prevision in $\mathcal{M}(\underline{P})$ which agrees with P on \mathcal{T} . The equivalent condition 3.3.2(a) reduces to the condition in the theorem, using the fact that P is linear on the linear space \mathcal{T} . Since $P \geq \underline{P}$ on \mathcal{T} , the stated condition is strengthened by replacing $P(X_0)$ by $\underline{P}(X_0)$. It is then equivalent to Definition 2.5.1 (for $X_0 \in \mathcal{T}$) since $-X_0 + \underline{P}(X_0) = -G(X_0)$. ♦

Provided \underline{P} is coherent on \mathcal{K} , any linear prevision that dominates \underline{P} on a linear subspace \mathcal{T} can be extended to \mathcal{K} as a dominating linear prevision.⁷ When \mathcal{K} is itself a linear space, the coherent \underline{P} are positively homogeneous, super-linear functionals (by Theorem 2.5.5), and the linear previsions P are linear functionals (by Corollary 2.8.5). In that case the theorem yields a version of the well-known Hahn–Banach extension theorem.

3.5 Invariant linear previsions

In this section we apply the fundamental results on avoiding sure loss, coherence and natural extension to characterize the linear previsions that are invariant under a semigroup of transformations. (The same approach can be used more generally to characterize the linear previsions that satisfy a given set of constraints.) As examples we will characterize the shift-invariant linear previsions on \mathbb{Z}^+ (known as Banach limits), the translation-invariant linear previsions on \mathbb{R} , and the translation-invariant extensions of Lebesgue measure to all subsets of $[0, 1]$.

3.5.1 Definitions

Throughout this section \mathcal{G} will denote a **semigroup of transformations** of Ω . That is, each $g \in \mathcal{G}$ maps Ω into itself, and the composition $g_1 g_2$, defined by $g_1 g_2(\omega) = g_1(g_2(\omega))$, is in \mathcal{G} whenever g_1 and g_2 are in \mathcal{G} .¹ The semigroup \mathcal{G} is **Abelian** if $g_1 g_2 = g_2 g_1$ whenever $g_1, g_2 \in \mathcal{G}$.

For $X \in \mathcal{L}$ and $g \in \mathcal{G}$, define $Xg \in \mathcal{L}$ by $(Xg)(\omega) = X(g(\omega))$. Thus g can be regarded as a mapping of \mathcal{L} into itself. This mapping is linear, i.e. $(X + Y)g = Xg + Yg$, homogeneous, i.e. $(\lambda X)g = \lambda(Xg)$ for $\lambda \in \mathbb{R}$, monotonic, i.e. if $X \geq Y$ then $Xg \geq Yg$, and fixes constant gambles, i.e. $\lambda g = \lambda$. Since $\{g(\omega) : \omega \in \Omega\}$ is a subset of Ω , the range of Xg is contained in the range of X and $\inf X \leq \inf(Xg) \leq \sup(Xg) \leq \sup X$. Also $X(g_1 g_2) = (Xg_1)g_2$ whenever $X \in \mathcal{L}$ and $g_1, g_2 \in \mathcal{G}$. If \mathcal{G} is Abelian then $Xg_1 g_2 = Xg_2 g_1$.

A lower revision \underline{P} on \mathcal{K} is called \mathcal{G} -invariant if $\underline{P}(Xg) = \underline{P}(X)$ whenever $g \in \mathcal{G}, X \in \mathcal{K}$ and $Xg \in \mathcal{K}$. We are especially interested in \mathcal{G} -invariant linear previsions, which are often called invariant means.

For example, let $\Omega = \mathbb{R}$ and define the translations g_c by $g_c(\omega) = \omega + c$. The Abelian group of transformations $\mathcal{G} = \{g_c; c \in \mathbb{R}\}$ is called the translation group on \mathbb{R} . Since $Xg_c(\omega) = X(\omega + c)$, \mathcal{G} -invariance is translation-invariance.

It is sometimes necessary to establish the existence of linear previsions that are \mathcal{G} -invariant and also satisfy further constraints. We start with two general results which characterize the \mathcal{G} -invariant linear previsions that dominate a given lower revision \underline{P} . A necessary and sufficient condition for the existence of such linear previsions is that \underline{P} avoids sure loss when it is extended by defining $\underline{P}(X - Xg) = 0$ for all $X \in \mathcal{L}$ and $g \in \mathcal{G}$. We will then look at the special cases where \underline{P} is vacuous, to characterize the class of all \mathcal{G} -invariant linear previsions, and where \mathcal{G} is Abelian, which guarantees that \mathcal{G} -invariant linear previsions exist.

3.5.2 Dominance theorem

Suppose that \mathcal{G} is a semigroup of transformations of Ω , the lower revision \underline{P} is defined on a domain \mathcal{K} that is closed under \mathcal{G} (i.e., if $X \in \mathcal{K}$ and $g \in \mathcal{G}$ then $Xg \in \mathcal{K}$), \underline{P} avoids sure loss, and $\underline{P}(Xg) \geq \underline{P}(X)$ whenever $X \in \mathcal{K}$ and $g \in \mathcal{G}$. (The last condition obviously holds if \underline{P} is \mathcal{G} -invariant.) Let \underline{E} denote the natural extension of \underline{P} to \mathcal{L} and \bar{E} its conjugate upper revision. Then there are \mathcal{G} -invariant linear previsions (defined on \mathcal{L}) which dominate \underline{P} on \mathcal{K} if and only if

$$\bar{E}\left(\sum_{j=1}^n (X_j - X_j g_j)\right) \geq 0 \quad \text{whenever } n \geq 1, X_j \in \mathcal{L} \quad \text{and} \quad g_j \in \mathcal{G}.$$

Proof. The key step in the proof, (iii) below, is to construct \underline{E}' which dominates \underline{E} so that $\mathcal{M}(\underline{E}')$ contains just the \mathcal{G} -invariant linear previsions in $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{P})$. Then $\mathcal{M}(\underline{E}')$ is non-empty if and only if \underline{E}' avoids sure loss, and this reduces to the condition stated in the theorem.

- (i) By 3.1.2 and 3.4.1, \underline{E} is coherent and $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{P})$.
- (ii) Use Definition 3.1.1, and the assumption that $\underline{P}(Xg) \geq \underline{P}(X)$ when $X \in \mathcal{K}$, to show that $\underline{E}(Xg) \geq \underline{E}(X)$ whenever $X \in \mathcal{L}$ and $g \in \mathcal{G}$.
- (iii) Let $\mathcal{T} = \{X - Xg: X \in \mathcal{L}, g \in \mathcal{G}\}$. Define \underline{E}' on \mathcal{L} by $\underline{E}'(Y) = 0$ for $Y \in \mathcal{T}$ and $\underline{E}'(Y) = \underline{E}(Y)$ for $Y \in \mathcal{L} - \mathcal{T}$. Note that $\mathcal{T} = -\mathcal{T}$ since $(-X) - (-X)g = -(X - Xg)$, so $\bar{E}'(Y) = -\underline{E}'(-Y) = 0$ for $Y \in \mathcal{T}$. Using coherence of \underline{E} and (ii), $\underline{E}(X - Xg) \leq \underline{E}(X) - \underline{E}(Xg) \leq 0$, so $\underline{E}(Y) \leq 0$ for $Y \in \mathcal{T}$. Hence $\underline{E}' \geq \underline{E}$ on \mathcal{L} and $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{E}')$.
- (iv) Show that $\mathcal{M}(\underline{E}')$ contains just the \mathcal{G} -invariant linear previsions in $\mathcal{M}(\underline{P})$.
- (v) There are such linear previsions if and only if \underline{E}' avoids sure loss, by

3.5 INVARIANT LINEAR PREVISIONS

3.3.3(a). By 2.4.1 and definition of \underline{E}' , that is equivalent to the condition: $\sup(U + V) \geq 0$ whenever $U = \sum_{j=1}^n (X_j - X_j g_j)$, $V = \sum_{i=1}^m (Z_i - \underline{E}(Z_i))$, $X_j \in \mathcal{L}, g_j \in \mathcal{G}$ and $Z_i \in \mathcal{L}$. This implies that $\bar{E}(U) \geq 0$ by taking $m = 1$ and $Z_1 = -U$. Conversely, if $\bar{E}(U) \geq 0$ then, using coherence of \underline{E} ,

$$\sup(U + V) \geq \bar{E}(U + V) \geq \bar{E}(U) + \underline{E}(V) \geq \underline{E}(V) \geq \sum_{i=1}^m \underline{E}(Z_i - \underline{E}(Z_i)) = 0. \quad \diamond$$

Of course, if \underline{P} is coherent on domain \mathcal{L} then its natural extension \underline{E} agrees with \underline{P} . Steps (i) and (ii) of the proof show that, under the assumptions of the theorem, there is no loss of generality in assuming that \underline{P} is coherent and $\mathcal{K} = \mathcal{L}$, since we can achieve that by replacing \underline{P} by \underline{E} while retaining the property $\underline{E}(Xg) \geq \underline{E}(X)$.²

Let $\mathcal{M}[\mathcal{G}, \underline{P}]$ denote the class of all linear previsions on \mathcal{L} which dominate \underline{P} on its domain \mathcal{K} and are \mathcal{G} -invariant. The next theorem gives several characterizations of this class.

3.5.3 Theorem

Under the assumptions of Theorem 3.5.2,

$$\mathcal{M}[\mathcal{G}, \underline{P}] = \mathcal{M}(\underline{E}') = \mathcal{M}(\underline{R}) = \mathcal{M}(\underline{Q}),$$

where \underline{E}' is the lower revision defined in the proof of 3.5.2, \underline{R} and \underline{Q} are defined on \mathcal{L} by

$$\underline{R}(X) = \sup \left\{ \underline{E}\left(X - \sum_{j=1}^n (X_j - X_j g_j)\right) : n \geq 0, X_j \in \mathcal{L}, g_j \in \mathcal{G} \right\},$$

and

$$\underline{Q}(X) = \sup \left\{ \underline{E}\left(n^{-1} \sum_{j=1}^n X_j g_j\right) : n \geq 1, g_j \in \mathcal{G} \right\}.$$

If $\mathcal{M}[\mathcal{G}, \underline{P}]$ is non-empty then \underline{R} is its lower envelope.

Proof. (i) $\mathcal{M}[\mathcal{G}, \underline{P}] = \mathcal{M}(\underline{E}')$ by (iv) of the previous proof.

(ii) Let \underline{R} be the natural extension of \underline{E}' , so $\mathcal{M}(\underline{R}) = \mathcal{M}(\underline{E}')$ by 3.4.1. Formula 3.1.1 for the natural extension may be written as

$$\underline{R}(X) = \sup\{\alpha: X - \alpha \geq U + V\},$$

where U and V are as in (v) of the previous proof. Use coherence of \underline{E}' to reduce this to the expression in the theorem. Provided \underline{E}' avoids sure loss, its natural extension \underline{R} is coherent and therefore the lower envelope of $\mathcal{M}(\underline{R})$.

(iii) It remains to verify that $\mathcal{M}[\mathcal{G}, \underline{P}] = \mathcal{M}(\underline{Q})$. Note first that $\underline{Q}(X)$ is finite, since

$$\underline{E}\left(n^{-1} \sum_{j=1}^n X_j g_j\right) \leq \sup\left(n^{-1} \sum_{j=1}^n X_j g_j\right) \leq n^{-1} \sum_{j=1}^n \sup(X_j g_j) \leq \sup X.$$

Show that $\underline{Q} \leq \underline{R}$ by taking $X_j = n^{-1}X$ in the formula for \underline{R} . This proves that $\mathcal{M}(\underline{Q}) \supset \mathcal{M}(\underline{R}) = \mathcal{M}[\mathcal{G}, \underline{P}]$.

(iv) Show that $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{P}) \supset \mathcal{M}(\underline{Q})$, by taking $n = 1$ in the formula for \underline{Q} to give $\underline{Q}(X) \geq \underline{E}(Xg) \geq \underline{E}(X)$, using (ii) of 3.5.2. To prove that all $P \in \mathcal{M}(\underline{Q})$ are \mathcal{G} -invariant, take $X = Y - Yg$ and $g_j = g^{j-1}$ to give $\underline{Q}(Y - Yg) \geq \underline{E}(n^{-1} \sum_{j=1}^n (Yg^{j-1} - Yg^j)) = n^{-1} \underline{E}(Y - Yg^n) \geq -2n^{-1} \sup |Y|$. Let $n \rightarrow \infty$ to give $\underline{Q}(Y - Yg) \geq 0$. By replacing Y by $-Y$, $\underline{Q}(Yg - Y) \geq 0$. Hence, for all $P \in \mathcal{M}(\underline{Q})$, $P(Yg - Y) = 0$ and P is \mathcal{G} -invariant. This proves that $\mathcal{M}[\mathcal{G}, \underline{P}] \supset \mathcal{M}(\underline{Q})$. ♦

Let $\mathcal{M}[\mathcal{G}]$ denote the class of all \mathcal{G} -invariant linear previsions on \mathcal{L} . By taking \underline{P} to be the vacuous lower prevision in the previous theorems, we obtain the following characterizations of $\mathcal{M}[\mathcal{G}]$.

3.5.4 Corollary

Let \mathcal{G} be a semigroup of transformations of Ω .

(a) $\mathcal{M}[\mathcal{G}]$ is non-empty if and only if

$$\sup \sum_{j=1}^n (X_j - X_j g_j) \geq 0 \quad \text{whenever } n \geq 1, X_j \in \mathcal{L} \quad \text{and} \quad g_j \in \mathcal{G}.^3$$

(b) $\mathcal{M}[\mathcal{G}] = \mathcal{M}(\underline{R}) = \mathcal{M}(\underline{Q})$, where \underline{R} and \underline{Q} are defined on \mathcal{L} by

$$\underline{R}(X) = \sup \left\{ \inf \left(X - \sum_{j=1}^n (X_j - X_j g_j) \right) : n \geq 0, X_j \in \mathcal{L}, g_j \in \mathcal{G} \right\},$$

and

$$\underline{Q}(X) = \sup \left\{ \inf \left(n^{-1} \sum_{j=1}^n X_j g_j \right) : n \geq 1, g_j \in \mathcal{G} \right\}.$$

Condition (a) is satisfied if and only if \underline{R} is coherent, and then \underline{R} is the lower envelope of $\mathcal{M}[\mathcal{G}]$.

Proof. Apply 3.5.2 and 3.5.3 with $\mathcal{K} = \mathcal{L}$ and \underline{P} vacuous, so that $\underline{P}(Xg) = \inf(Xg) \geq \inf X = \underline{P}(X)$, and \underline{E} is also vacuous. ♦

It can be difficult to verify the general conditions for the existence of \mathcal{G} -invariant linear previsions, 3.5.2 and 3.5.4(a), but we can show that they are always satisfied when the semigroup \mathcal{G} is Abelian. We do so by proving that the lower prevision \underline{Q} defined in 3.5.3 is coherent.

3.5.5 Abelian dominance theorem⁴

Suppose that the assumptions of 3.5.2 hold, and also \mathcal{G} is Abelian. Then $\mathcal{M}[\mathcal{G}, \underline{P}]$ is non-empty, the lower previsions \underline{R} and \underline{Q} defined in 3.5.3 agree, and each is the lower envelope of $\mathcal{M}[\mathcal{G}, \underline{P}]$.

3.5 INVARIANT LINEAR PREVISIONS

Proof. Verify that \underline{Q} satisfies the coherence axioms P1–P3 of Definition 2.3.3. We use throughout the fact that \underline{E} is coherent. P1 holds since $\underline{Q}(X) \geq \underline{E}(Xg) \geq \inf X$. P2 holds because $(\lambda X)g = \lambda(Xg)$ and $\underline{E}(\lambda Y) = \lambda \underline{E}(Y)$.

For P3, suppose $X, Y \in \mathcal{L}$. We need to prove $\underline{Q}(X + Y) \geq \underline{Q}(X) + \underline{Q}(Y)$. Given $\delta > 0$, find $U = n^{-1} \sum_{j=1}^n X_j g_j$ and $V = m^{-1} \sum_{i=1}^m Y_i h_i$ such that $\underline{E}(U) \geq \underline{Q}(X) - \delta$ and $\underline{E}(V) \geq \underline{Q}(Y) - \delta$. Using the assumption that \mathcal{G} is Abelian,

$$W = m^{-1} n^{-1} \sum_{i=1}^m \sum_{j=1}^n (X_i + Y_i) h_i g_j = m^{-1} \sum_{i=1}^m U h_i + n^{-1} \sum_{j=1}^n V g_j.$$

Hence $\underline{Q}(X + Y) \geq \underline{E}(W) \geq m^{-1} \sum_{i=1}^m \underline{E}(U h_i) + n^{-1} \sum_{j=1}^n \underline{E}(V g_j) \geq \underline{E}(U) + \underline{E}(V) \geq \underline{Q}(X) + \underline{Q}(Y) - 2\delta$, using (ii) of Theorem 3.5.2. Since δ can be arbitrarily small, this verifies P3.

Since \underline{Q} is coherent, $\mathcal{M}(\underline{Q})$ is non-empty and \underline{Q} is its lower envelope by Theorem 3.3.3. The result then follows from Theorem 3.5.3. ♦

3.5.6 Corollary⁵

Let \mathcal{G} be an Abelian semigroup of transformations of Ω . Then $\mathcal{M}[\mathcal{G}]$ is non-empty and $\mathcal{M}[\mathcal{G}] = \mathcal{M}(\underline{Q})$, where \underline{Q} is the coherent lower prevision defined on \mathcal{L} by $\underline{Q}(X) = \sup \{\inf(n^{-1} \sum_{j=1}^n X_j g_j) : n \geq 1, g_j \in \mathcal{G}\}$.

Proof. Apply Theorem 3.5.5 with $\mathcal{K} = \mathcal{L}$ and \underline{P} vacuous. ♦

These results will now be applied to three of the examples introduced in section 2.9, to characterize the translation- or shift-invariant linear previsions.

3.5.7 Banach limits

Here $\Omega = \mathbb{Z}^+$. Let g_m denote the right shift by m , $g_m(\omega) = \omega + m$, and $\mathcal{G} = \{g_m : m \in \mathbb{Z}^+\}$. Then \mathcal{G} is an Abelian semigroup, since the composition of the shifts by m and by n is the shift by $m+n$, irrespective of order.⁶ By Corollary 3.5.6 there are shift-invariant linear previsions on $\mathcal{L}(\mathbb{Z}^+)$. These are called **Banach limits**.⁷ The class of all Banach limits has lower envelope $\underline{Q}(X) = \sup \{\inf_{\omega \in \mathbb{Z}^+} n^{-1} \sum_{j=1}^n X(\omega + m_j) : n, m_j \in \mathbb{Z}^+\}$, again by Corollary 3.5.6. It is a straightforward exercise to show that this can be written in the alternative form noted in section 2.9.5, $\underline{Q}(X) = \lim_{n \rightarrow \infty} \inf_{k \geq 1} n^{-1} \sum_{j=k}^{k+n-1} X(j)$. The upper envelope of all Banach limits is the conjugate upper prevision \bar{Q} , obtained by replacing ‘inf’ by ‘sup’ in the last formula.

Gambles X in $\mathcal{L}(\mathbb{Z}^+)$ can be regarded as bounded sequences of real numbers $X(n)$. A Banach limit P defines a sort of ‘generalized limit’ of the sequence X . Whenever $X(n)$ converges to x , every Banach limit $P \in \mathcal{M}(\underline{Q})$

must satisfy $P(X) = x$, because $n^{-1} \sum_{j=k}^{k+n-1} X(j) \rightarrow x$ (uniformly in k) as $n \rightarrow \infty$, so that $\underline{Q}(X) = \bar{Q}(X) = x$. More generally, all Banach limits P satisfy $\liminf_{n \rightarrow \infty} X(n) \leq \underline{Q}(X) \leq P(X) \leq \bar{Q}(X) \leq \limsup_{n \rightarrow \infty} X(n)$ for all $X \in \mathcal{L}$.⁸

A smaller class of Banach limits can be obtained as the class $\mathcal{M}(\underline{P})$ for the lower prevision \underline{P} defined in section 2.9.5, $\underline{P}(X) = \liminf_{n \rightarrow \infty} n^{-1} \sum_{j=1}^n X(j)$. Thus for all $X \in \mathcal{L}$, $n^{-1} \sum_{j=1}^n (X - Xg_1)(j) = n^{-1} \sum_{j=1}^n (X(j) - X(j+1)) = n^{-1} (X(1) - X(n+1)) \rightarrow 0$ as $n \rightarrow \infty$, so that $\underline{P}(X - Xg_1) = \bar{P}(X - Xg_1) = 0$. Hence all linear previsions in $\mathcal{M}(\underline{P})$ are Banach limits. Since \underline{P} is coherent, it follows directly from Theorem 3.3.3 that $\mathcal{M}(\underline{P})$ is non-empty and Banach limits exist (without appealing to the general result 3.5.6). The class $\mathcal{M}(\underline{P})$ is considerably smaller than $\mathcal{M}(\underline{Q})$ as there are sets A with $\underline{P}(A) = \bar{P}(A) = 0$ but $\underline{Q}(A) = 1$, e.g. let $A = \{2^n + 1, \dots, 2^n + n : n \in \mathbb{Z}^+\}$.

Notice that whereas \underline{P} and \underline{Q} are explicitly defined as lower limits, the Banach limits in $\mathcal{M}(\underline{P})$ or $\mathcal{M}(\underline{Q})$ cannot be explicitly defined or constructed.

3.5.8 Translation-invariant linear previsions on \mathbb{R}

Consider next the translation group $\mathcal{G} = \{g_c : c \in \mathbb{R}\}$ on $\Omega = \mathbb{R}$, where $g_c(\omega) = \omega + c$. Applying Corollary 3.5.6, there are translation-invariant linear previsions on $\mathcal{L}(\mathbb{R})$ and their lower envelope is $\underline{Q}(X) = \sup\{\inf_{\omega \in \mathbb{R}} n^{-1} \sum_{j=1}^n X(\omega + c_j) : n \geq 1, c_j \in \mathbb{R}\}$.

Compare \underline{Q} with the lower previsions \underline{P} and \underline{Q}' defined in section 2.9.7, $\underline{P}(X) = \liminf_{c \rightarrow \infty} (1/2c) \int_{-\infty}^c X(t) dt$ and $\underline{Q}'(X) = \lim_{c \rightarrow \infty} \inf_{x \in \mathbb{R}} (1/2c) \int_{x-c}^{x+c} X(t) dt$. It might be expected, by analogy with Banach limits, that $\underline{Q} = \underline{Q}'$. In fact $\underline{Q}(X) = \underline{Q}'(X)$ whenever X is uniformly continuous on \mathbb{R} (since then the integral can be approximated by a finite sum, uniformly in x), but $\underline{Q}'(X) > \underline{Q}(X)$ for some gambles X . There are Borel sets A for which $\underline{Q}'(A) = 1$ but $\underline{Q}(A) = 0$, so $P(A) = 0$ for some translation-invariant P .⁹ In general we have $\bar{P} \geq \underline{Q}' \geq \underline{Q}$, with strict inequality for some gambles. All the linear previsions in the three classes $\mathcal{M}(\underline{Q}) \supset \mathcal{M}(\underline{Q}') \supset \mathcal{M}(\underline{P})$ are translation-invariant.¹⁰

3.5.9 Lebesgue measure

As a final example we characterize the translation-invariant linear previsions that extend Lebesgue measure to all gambles on $\Omega = [0, 1]$. As in section 2.9.6, let v be Lebesgue measure, defined on the Borel σ -field B of Ω . Let $g_c(\omega) = \omega \oplus c$ be translation modulo 1, and $\mathcal{G} = \{g_c : 0 \leq c < 1\}$, an Abelian group. Note that B is closed under \mathcal{G} and v is a translation-invariant additive probability on B . Applying Theorem 3.5.5, there are translation-invariant linear previsions in $\mathcal{M}(v)$. Since $\mathcal{M}(v)$ is just the class of all linear extensions of v to \mathcal{L} (by Theorem 3.4.2), this proves that there are translation-invariant extensions of Lebesgue measure to all gambles. (Of course, none of these is countably additive.)

3.6 COMPACTNESS AND EXTREME POINTS OF $\mathcal{M}(\underline{P})$

Let $X \oplus c = Xg_c$ denote the translate of X by c . Also by Theorem 3.5.5, the lower envelope of all translation-invariant linear extensions is

$$\begin{aligned} \underline{Q}(X) &= \sup \left\{ \underline{E} \left(n^{-1} \sum_{j=1}^n X \oplus c_j \right) : n \geq 1, 0 \leq c_j < 1 \right\} \\ &= \sup \left\{ P_0(Y) : Y \in \mathcal{K}, Y \leq n^{-1} \sum_{j=1}^n X \oplus c_j, n \geq 1, 0 \leq c_j < 1 \right\}, \end{aligned}$$

where P_0 is the unique linear extension of v to the linear space \mathcal{K} of Borel-measurable gambles (defined by Theorem 3.2.2), and \underline{E} is the natural extension of v to \mathcal{L} , defined by $\underline{E}(X) = \sup \{P_0(Y) : Y \in \mathcal{K}, Y \leq X\}$. For some non-measurable gambles X , $\underline{E}(X)$ is smaller than $\underline{Q}(X)$, which means that there are finitely additive extensions of Lebesgue measure to all subsets of Ω that are not translation-invariant.¹¹

3.6 Compactness and extreme points of $\mathcal{M}(\underline{P})$

We now introduce a topology on the space \mathcal{P} of linear previsions. Throughout this section \mathcal{P} is given the **weak* topology**, under which it is **compact**. (The weak* topology is defined in Appendix D, which also contains the necessary background material on linear topological spaces.) This enables us to characterize the subsets of \mathcal{P} that arise as classes of dominating linear previsions $\mathcal{M}(\underline{P})$, and to show that every coherent lower prevision \underline{P} is the lower envelope of the set of extreme points of $\mathcal{M}(\underline{P})$.¹ By applying these general results to the example of 0–1 valued lower probabilities, we will prove the ultrafilter theorem and hence establish the existence of non-degenerate ultrafilters and non-measurable sets.

Suppose that the lower prevision $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss, so $\mathcal{M}(\underline{P})$ is non-empty. Regard $\mathcal{M}(\underline{P})$ as the intersection over all $X \in \mathcal{K}$ of the sets $\{P \in \mathcal{P} : P(X) \geq \underline{P}(X)\}$. Each of these sets is convex and closed, hence compact, under the weak* topology. Their intersection $\mathcal{M}(\underline{P})$ is therefore convex and compact.² In fact there is the following one-to-one correspondence between convex compact classes of linear previsions and coherent lower previsions on \mathcal{L} .

3.6.1 Weak*-compactness theorem³

There is a one-to-one correspondence between the coherent lower previsions (\underline{P}) on domain \mathcal{L} and the non-empty weak*-compact convex subsets (\mathcal{M}) of \mathcal{P} : \underline{P} is the lower envelope of its corresponding \mathcal{M} , and \mathcal{M} is the set $\mathcal{M}(\underline{P})$ of all linear previsions that dominate \underline{P} .

Proof. Consider the mapping $\underline{P} \mapsto \mathcal{M}(\underline{P})$. We have shown that $\mathcal{M}(\underline{P})$ is

convex and compact. The lower envelope of $\mathcal{M}(\underline{P})$ is \underline{P} by 3.3.3, so the mapping is one-to-one. Let \mathcal{M} be an arbitrary non-empty compact convex subset of \mathcal{P} . Let \underline{P} be the lower envelope of \mathcal{M} , coherent by Theorem 3.3.3. To show that the mapping is bijective we need to show that $\mathcal{M} = \mathcal{M}(\underline{P})$. We have $\mathcal{M}(\underline{P}) \supset \mathcal{M}$ since \mathcal{M} has lower envelope \underline{P} . To show that $\mathcal{M} \supset \mathcal{M}(\underline{P})$, suppose $Q \in \mathcal{P} - \mathcal{M}$. By the strong separation theorem (Appendix E4) there is a continuous linear functional Λ and $\alpha \in \mathbb{R}$ such that $\Lambda(Q) < \alpha$ and $\Lambda(P) \geq \alpha$ for all $P \in \mathcal{M}$. But the weak*-continuous linear functionals are just the evaluation functionals X^* (Appendix D3), so $\Lambda(P) = P(X)$ for some $X \in \mathcal{L}$. Thus $P(X) = \min\{P(X): P \in \mathcal{M}\} \geq \alpha$ but $Q(X) < \alpha$. This proves that Q is not in $\mathcal{M}(\underline{P})$. Thus $\mathcal{M} \supset \mathcal{M}(\underline{P})$ as required. ♦

So all the weak*-compact convex subsets of \mathcal{P} (and only these subsets) can be obtained as classes $\mathcal{M}(\underline{P})$ for some lower prevision \underline{P} .⁴ If \underline{P} is coherent on \mathcal{L} then there may be many classes \mathcal{M} which have \underline{P} as their lower envelope, but $\mathcal{M}(\underline{P})$ is the only compact convex class which does.

Theorem 3.6.1 can be used to give a simple proof of the natural extension theorem 3.4.1. Suppose that $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss. By property 3.1.2(e), the natural extension \underline{E} is the minimal coherent lower prevision on \mathcal{L} that dominates \underline{P} on \mathcal{K} . Using the correspondence 3.6.1, $\mathcal{M}(\underline{E})$ is the maximal compact convex set contained in $\mathcal{M}(\underline{P})$. But $\mathcal{M}(\underline{P})$ is itself a non-empty compact convex set, so $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{P})$.

The proof of Theorem 3.6.1 relies on the fact that any closed convex set \mathcal{M} in a locally convex topological space is the intersection of the closed ‘half-spaces’ containing it. Under the weak* topology, \mathcal{M} is the intersection of the closed half-spaces $\{P \in \mathcal{P}: P(X) \geq \underline{P}(X)\}$ over all $X \in \mathcal{L}$, where \underline{P} is the lower envelope of \mathcal{M} . Each gamble X defines a **supporting hyperplane** $\{P \in \mathcal{P}: P(X) = \underline{P}(X)\}$ for \mathcal{M} . We say that X supports \mathcal{M} at those points of \mathcal{M} which lie on the hyperplane.

Every X supports $\mathcal{M}(\underline{P})$ at some point, since $\underline{P}(X)$ is attained by some $P \in \mathcal{M}(\underline{P})$. The following theorem shows that every gamble supports $\mathcal{M}(\underline{P})$ at some **extreme point**, that is, an element of $\mathcal{M}(\underline{P})$ which cannot be written as a convex combination of other elements. Hence, a coherent lower prevision \underline{P} is the lower envelope of the set of extreme points of $\mathcal{M}(\underline{P})$. The theorem is a direct corollary of the Krein–Milman theorem (Appendix E5).

3.6.2 Extreme point theorem

Suppose the lower prevision $(\Omega, \mathcal{K}, \underline{P})$ avoids sure loss. Let $\text{ext } \mathcal{M}(\underline{P})$ denote the set of all extreme points of $\mathcal{M}(\underline{P})$.

- (a) $\text{ext } \mathcal{M}(\underline{P})$ is non-empty.
- (b) $\mathcal{M}(\underline{P})$ is the weak*-closure of the convex hull of $\text{ext } \mathcal{M}(\underline{P})$, i.e. the smallest weak*-compact convex set containing $\text{ext } \mathcal{M}(\underline{P})$.⁵

- (c) If \underline{P} is coherent then for every $X \in \mathcal{K}$ there is $P \in \text{ext } \mathcal{M}(\underline{P})$ such that $P(X) = \underline{P}(X)$. Thus \underline{P} is the lower envelope of $\text{ext } \mathcal{M}(\underline{P})$.⁶

From (c), in order to define a coherent lower prevision \underline{P} it suffices to specify the extreme points of $\mathcal{M}(\underline{P})$.

3.6.3 Examples

Here we identify the set of extreme points $\text{ext } \mathcal{M}(\underline{P})$ for some of the lower previsions \underline{P} defined in section 2.9. (The classes $\mathcal{M}(\underline{P})$ were identified in section 3.3.5.)

(a) Vacuous previsions

Here $\mathcal{M}(\underline{P}) = \mathcal{P}$. It is shown in 3.6.7 that the extreme points of \mathcal{P} are just the expectations under 0–1 valued additive probabilities.

(b) Linear–vacuous mixtures

Here $\text{ext } \mathcal{M}(\underline{P}) = \{(1 - \delta)P_0 + \delta P: P \in \text{ext } \mathcal{P}\}$.

(c) Pari-mutuel models

If Ω is finite, each ordering of Ω corresponds to an extreme point P determined by

$$P(\{\omega_j\}) = (1 + \delta)P_0(\{\omega_j\}) \quad \text{for } 1 \leq j < n, \quad \text{and} \quad P(\{\omega_j\}) = 0 \quad \text{for } j > n.$$

(d) Constant odds-ratio models

Each non-empty set A corresponds to an extreme point P_A defined by

$$P_A(X) = \frac{(1 - \tau)P_0(AX) + P_0(A^c X)}{1 - \tau P_0(A)}.$$

(P_A is obtained from P_0 by reducing the probabilities of points in A by a factor $1 - \tau$, then renormalizing.)

(e) Zero–one valued probabilities

Recall the correspondence (2.9.8) between coherent 0–1 valued lower probabilities \underline{P} and filters \mathcal{A} . The next theorem identifies $\text{ext } \mathcal{M}(\underline{P})$ for 0–1 valued \underline{P} . The rest of this section develops the relationship between 0–1 valued probabilities, filters and ultrafilters.

Let \mathcal{Q} denote the class of linear previsions on \mathcal{L} whose restrictions to events take only the values 0 and 1.⁷ Because linear previsions on \mathcal{L} are uniquely determined by their restrictions to events (Corollary 3.2.3), we can identify \mathcal{Q} with the class of 0–1 valued additive probabilities defined on all subsets of Ω . (These correspond to ultrafilters of subsets, by the results in section 2.9.8.)

3.6.4 Theorem

Let \underline{P} be the 0–1 valued lower probability corresponding to the filter \mathcal{A} , defined for all events A by $\underline{P}(A) = 1$ if $A \in \mathcal{A}$, otherwise $\underline{P}(A) = 0$. Then $\text{ext } \mathcal{M}(\underline{P}) = \{\underline{P} \in \mathcal{Q}: \underline{P}(A) = 1 \text{ for all } A \in \mathcal{A}\}$. These extreme points correspond to the ultrafilters containing \mathcal{A} .

Proof. Suppose $P \in \mathcal{Q}$ and $P(A) = 1$ for all $A \in \mathcal{A}$. Then clearly $P \in \mathcal{M}(\underline{P})$. If $P = \lambda P_1 + (1 - \lambda)P_2$ for some $P_1, P_2 \in \mathcal{P}$ and $0 < \lambda < 1$, then P_1 and P_2 must agree with P on events since P is 0–1 valued. Using Corollary 3.2.3, $P \in \text{ext } \mathcal{M}(\underline{P})$.

Conversely, suppose $P \in \mathcal{M}(\underline{P})$ but $0 < P(B) < 1$ for some event B . Write $\beta = P(B)$. Define P_1 and P_2 for $X \in \mathcal{L}$ by

$$P_1(X) = (1 - \beta)P(X) + P(BX) \quad \text{and} \quad P_2(X) = (1 + \beta)P(X) - P(BX).$$

Then P_1 and P_2 are linear functionals, and they satisfy axiom P1 since

$$P_1(X) \geq (1 - \beta)\inf X + P(B)\inf X = \inf X$$

and

$$P_2(X) = P((1 + \beta - B)X) \geq P(1 + \beta - B)\inf X = \inf X.$$

By Theorem 2.8.4, P_1 and P_2 are linear previsions on \mathcal{L} . When $A \in \mathcal{A}$ we have $P(A) = 1$, hence $P(BA) = P(B) = \beta$ and $P_1(A) = P_2(A) = 1$. Thus P_1 and P_2 are in $\mathcal{M}(\underline{P})$. But $P = \frac{1}{2}P_1 + \frac{1}{2}P_2$, and $P_1 \neq P$ because $P_1(B) = (2 - \beta)\beta > \beta = P(B)$, so P is not an extreme point of $\mathcal{M}(\underline{P})$. This proves that $\mathcal{Q} \supset \text{ext } \mathcal{M}(\underline{P})$. The result follows by noting that $\mathcal{M}(\underline{P}) = \{P \in \mathcal{P}: P(A) = 1 \text{ for all } A \in \mathcal{A}\}$. ◆

The theorem does not establish that $\text{ext } \mathcal{M}(\underline{P})$ is non-empty or (equivalently) that there are ultrafilters containing \mathcal{A} . To prove that we must invoke the extreme point theorem and thereby establish the well-known result that every filter of sets can be extended to an ultrafilter.

3.6.5 Ultrafilter theorem⁸

If \mathcal{A} is a filter of subsets of Ω then there is an ultrafilter containing \mathcal{A} . Equivalently, there is $P \in \mathcal{Q}$ such that $P(A) = 1$ for all $A \in \mathcal{A}$.

Proof. Let \underline{P} be the coherent lower probability corresponding to \mathcal{A} . By Theorem 3.6.2(a), $\mathcal{M}(\underline{P})$ has extreme points. Apply Theorem 3.6.4. ◆

The ultrafilter theorem can be strengthened as follows, using the result that a coherent \underline{P} is the lower envelope of $\text{ext } \mathcal{M}(\underline{P})$.

3.6.6 Theorem

Let \underline{P} be a 0–1 valued lower probability defined on all subsets of Ω . Then \underline{P} avoids sure loss if and only if it is dominated by some $P \in \mathcal{Q}$, and \underline{P} is coherent if and only if it is the lower envelope of a subset of \mathcal{Q} . Hence, every filter is the intersection of all ultrafilters that contain it.

Proof. If \underline{P} avoids sure loss then $\text{ext } \mathcal{M}(\underline{P})$ is non-empty by Theorem 3.6.2(a). If \underline{P} is coherent then it is the lower envelope of $\text{ext } \mathcal{M}(\underline{P})$ by 3.6.2(c). The result then follows from Theorems 3.3.3 and 3.6.4. If \mathcal{A} is a filter, the corresponding \underline{P} is coherent. So \underline{P} can be written as a lower envelope of 0–1 valued additive probabilities, which corresponds to writing \mathcal{A} as an intersection of ultrafilters. ★

By applying Theorem 3.6.4 to the trivial filter $\mathcal{A} = \{\Omega\}$, which corresponds to the vacuous lower probability, we can characterize the extreme points of \mathcal{P} , the class of all linear previsions, as the 0–1 valued additive probabilities.

3.6.7 Corollary⁹

$\text{ext } \mathcal{P} = \mathcal{Q}$, and \mathcal{P} is the weak*-closure of the convex hull of \mathcal{Q} .

The set of extreme points \mathcal{Q} contains all the **degenerate** linear previsions, of the form $P(X) = X(\omega)$ for some $\omega \in \Omega$. The corresponding ultrafilters $\mathcal{A} = \{A: \omega \in A\}$ are also called degenerate.¹⁰ When Ω is finite, all ultrafilters are degenerate and $\text{ext } \mathcal{P} = \mathcal{Q}$ is just the set of degenerate linear previsions.¹¹

When Ω is infinite, however, there are non-degenerate extreme points and non-degenerate ultrafilters. To see that, define the filter $\mathcal{A} = \{A: A^c \text{ is finite}\}$, and use the ultrafilter theorem to show that there are ultrafilters containing \mathcal{A} . These cannot contain any finite set A since $A^c \in \mathcal{A}$, thus they are non-degenerate. The corresponding extreme points of \mathcal{P} satisfy $P(A) = 0$ for every finite set A . Thus we obtain the following result.

3.6.8 Weak ultrafilter theorem

There are non-degenerate ultrafilters of subsets of any infinite set Ω , i.e., there are $P \in \mathcal{Q}$ such that $P(A) = 0$ for all finite sets A .

The special case in which Ω is the set of positive integers is all that is required to prove the existence of sets that are not Lebesgue measurable.¹²

3.6.9 Non-measurable sets

Let \underline{P} and \bar{P} be inner and outer Lebesgue measure, defined on all subsets of $\Omega = [0, 1]$ as in section 2.9.6. There are subsets A of Ω for which $\bar{P}(A) > \underline{P}(A)$,

i.e., which are not Lebesgue measurable. In fact there are sets A with $\bar{P}(A) = 1$ and $\underline{P}(A) = 0$.

*Proof.*¹³ Let B be the set of binary rationals in $[0, 1]$. Every $x \in B^c$ has a unique binary expansion $x = \sum_{n=1}^{\infty} d_n(x)2^{-n}$, where $d_n(x)$ takes the values 0 and 1. Associate with x the set of positive integers $D(x) = \{n \in \mathbb{Z}^+ : d_n(x) = 1\}$. By the weak ultrafilter theorem there is $Q \in \mathcal{D}$, defined for all subsets D of \mathbb{Z}^+ , such that $Q(D) = 0$ whenever D is finite. Define a set $A = \{x \in B^c : Q(D(x)) = 1\}$. Let $C = (A \cup B)^c$, so $\{A, B, C\}$ is a partition of Ω . We show that A and C are non-measurable.

First verify that these sets have the properties:

- (a) $x \in A$ if and only if $1 - x \in C$,
- (b) $A \oplus z = A$ for all $z \in B$, where \oplus is translation modulo 1.¹⁴

For (a), note that $D(1 - x) = \{n : d_n(x) = 0\} = D(x)^c$, so $Q(D(x)) = 1$ just when $Q(D(1 - x)) = 0$. For (b), note that $D(x \oplus z) \triangle D(x)$ is finite, since the binary expansion of $x \oplus z$ agrees with the binary expansion of x except for finitely many digits. Using the fact that $Q(D) = 0$ when D is finite, $Q(D(x \oplus z)) = Q(D(x))$, hence $x \oplus z \in A$ just when $x \in A$.

By (a), the sets A and C are congruent and so $\bar{P}(A) = \bar{P}(C)$. Since B is a countable set, $\bar{P}(B) = 0$. Hence $1 = \bar{P}(\Omega) \leq \bar{P}(A) + \bar{P}(B) + \bar{P}(C) = 2\bar{P}(A)$, so $\bar{P}(A) \geq \frac{1}{2}$. Because B is dense in Ω , it follows from (b) that $\bar{P}(A) = 0$ or $\bar{P}(A) = 1$.¹⁵ Thus $\bar{P}(A) = \bar{P}(C) = 1$, and $\underline{P}(A) = 1 - \bar{P}(B \cup C) = 1 - \bar{P}(C) = 0$.

3.7 Desirability and preference

We now introduce some alternative ways of modelling beliefs, in terms of class of desirable gambles or a partial preference ordering of gambles. When they satisfy appropriate coherence conditions, these new models are mathematically equivalent to the previous models, coherent lower previsions or classes of linear previsions. However, the new models can contain more information than lower previsions as they can distinguish strict desirability or preference from almost-desirability or preference. We will present axioms for both strict and almost-desirability, as both are important in practice. Strict preference (or strict desirability) is the important notion when lower previsions are used in decision making (section 3.9), whereas almost-desirability (or almost-preference) is involved in elicitation of lower previsions (Chapter 4).

First we will define notions of avoiding sure loss and coherence for classes of almost-desirable gambles. (These correspond to the general definitions for lower previsions in Chapter 2.) After discussing almost-desirability in some detail we will outline the closely related models for almost-preference, strict desirability and strict preference.

Throughout, \mathcal{D} will denote a class of almost-desirable gambles. The behavioural interpretation of almost-desirability is that, for each X in \mathcal{D} and positive δ , You are disposed to accept the gamble $X + \delta$. (We say only that X is almost-desirable as You are not necessarily disposed to accept X itself.) It follows from this interpretation that we can expect \mathcal{D} to be closed in the supremum-norm topology: if $X + \delta \in \mathcal{D}$ for all $\delta > 0$ then $X \in \mathcal{D}$. The first definition, of avoiding sure loss, is a simpler version of the definition for lower previsions (2.4.1).

3.7.1 Definition

Suppose \mathcal{D} (a subset of \mathcal{L}) is regarded as a class of almost-desirable gambles. Then \mathcal{D} avoids sure loss if $\sup \sum_{j=1}^n X_j \geq 0$ whenever $n \geq 1$ and $X_1, \dots, X_n \in \mathcal{D}$.¹

In order to define coherence of \mathcal{D} it is convenient to first define its natural extension \mathcal{E} , which is the class of all gambles whose almost-desirability is effectively implied by the almost-desirability of gambles in \mathcal{D} . These are just the gambles which can be approximated by finite sums of gambles that are either in \mathcal{D} or non-negative.

3.7.2 Definition

Suppose $\mathcal{L} \supset \mathcal{K} \supset \mathcal{D}$, where \mathcal{K} is interpreted as a class of gambles whose desirability You have evaluated and \mathcal{D} as the subset of gambles judged almost desirable. Define \mathcal{E} , the natural extension of \mathcal{D} , to be the closure (in the supremum-norm topology) of

$$\left\{ Z \in \mathcal{L} : Z \geq \sum_{j=1}^n \lambda_j X_j \quad \text{for some } n \geq 0, \lambda_j > 0, X_j \in \mathcal{D} \right\},$$

which is the smallest convex cone containing \mathcal{D} and all non-negative gambles.² Note that $\mathcal{K} \cap \mathcal{E} \supset \mathcal{D}$. Say that \mathcal{D} is coherent relative to \mathcal{K} if \mathcal{D} avoids sure loss and also $\mathcal{D} = \mathcal{K} \cap \mathcal{E}$.

If the gambles in \mathcal{D} are almost-desirable then so should be the gambles in its natural extension \mathcal{E} . Your evaluations of the gambles in \mathcal{K} are coherent when \mathcal{D} avoids sure loss and contains all gambles in \mathcal{K} that are also in \mathcal{E} . Coherence of \mathcal{D} depends on the class \mathcal{K} in which \mathcal{D} is embedded, whereas avoiding sure loss is a property of \mathcal{D} alone.³

As expected, the definition of coherence simplifies when \mathcal{K} has suitable structure, especially when \mathcal{K} is a linear space. The following coherence axioms are just the axioms given in section 2.2.3 plus an extra closure axiom D4.

3.7.3 Axioms for almost-desirability

Suppose \mathcal{K} is a linear space containing constant gambles and $\mathcal{K} \supset \mathcal{D}$. Then \mathcal{D} is coherent relative to \mathcal{K} if and only if it satisfies the following five axioms:⁴

- (D0) if $X \in \mathcal{D}$ then $\sup X \geq 0$ (avoiding sure loss)
- (D1) if $X \in \mathcal{K}$ and $\inf X > 0$ then $X \in \mathcal{D}$ (accepting sure gains)
- (D2) if $X \in \mathcal{D}$ and $\lambda > 0$ then $\lambda X \in \mathcal{D}$ (positive homogeneity)
- (D3) if $X \in \mathcal{D}$ and $Y \in \mathcal{D}$ then $X + Y \in \mathcal{D}$ (addition)
- (D4) if $X \in \mathcal{K}$ and $X + \delta \in \mathcal{D}$ for all $\delta > 0$ then $X \in \mathcal{D}$ (closure).

Proof. Suppose \mathcal{D} satisfies the axioms. If $X_1, \dots, X_n \in \mathcal{D}$ then $\sum_{j=1}^n X_j \in \mathcal{D}$ by D3, and \mathcal{D} avoids sure loss by D0. By D1 and D4, \mathcal{D} contains all the non-negative gambles in \mathcal{K} . Suppose that $Z \in \mathcal{K}$ and $Z \geq \sum_{j=1}^n \lambda_j X_j$ for some $n \geq 1$, $\lambda_j > 0$, $X_j \in \mathcal{D}$. Then $\sum_{j=1}^n \lambda_j X_j \in \mathcal{D}$ by D2 and D3, hence $Z \in \mathcal{D}$ by D3 since $Z - \sum_{j=1}^n \lambda_j X_j$ is non-negative. Thus $\mathcal{D} \supset \mathcal{K} \cap \mathcal{V}$ where \mathcal{V} is the convex cone generated by \mathcal{D} and all non-negative gambles. If $X \in \mathcal{K} \cap \mathcal{E}$ then, for each $\delta > 0$, there is $Y \in \mathcal{V}$ such that $Y \leq X + \delta$, so $X + \delta \in \mathcal{K} \cap \mathcal{V}$ and $X + \delta \in \mathcal{D}$. Applying D4, $X \in \mathcal{D}$. Thus $\mathcal{D} \supset \mathcal{K} \cap \mathcal{E}$. Clearly $\mathcal{K} \cap \mathcal{E} \supset \mathcal{D}$. So \mathcal{D} is coherent relative to \mathcal{K} .

Conversely, suppose \mathcal{D} is coherent relative to \mathcal{K} . Then \mathcal{D} avoids sure loss, so D0 holds. The other axioms follow from $\mathcal{D} \supset \mathcal{K} \cap \mathcal{E}$ and the definition of \mathcal{E} . ♦

In particular, axioms D0–D4 characterize the classes of almost-desirable gambles that are coherent relative to \mathcal{L} . These are just the closed convex cones containing all non-negative gambles but not containing uniformly negative gambles. Axioms D0–D3 were justified in section 2.2.4. The closure axiom D4 is justified by our interpretation of \mathcal{D} as a class of almost-desirable, rather than really desirable, gambles. For example, the surely negative gamble in Example 2.4.5 is almost-desirable but not really desirable. The main reason for adopting this interpretation, and hence D4, is to establish the one-to-one correspondence between coherent classes \mathcal{D} and coherent lower previsions \underline{P} in Theorem 3.8.1, given by

$$\underline{P}(X) = \max \{\mu : X - \mu \in \mathcal{D}\} \quad \text{and} \quad \mathcal{D} = \{X \in \mathcal{K} : \underline{P}(X) \geq 0\}.$$

In eliciting \underline{P} , a judgement that X is really desirable tells us only that $\underline{P}(X) \geq 0$, which is no more informative about \underline{P} than a judgement that X is almost-desirable. So elicitation is concerned, in general, with judgements of almost-desirability.

Whenever \mathcal{D} avoids sure loss, its natural extension \mathcal{E} is coherent relative to \mathcal{L} . The following properties of \mathcal{E} correspond to the properties of the

3.7 DESIRABILITY AND PREFERENCE

natural extension \underline{E} given in section 3.1.2.⁵ (These properties can be verified using the axioms above.)

3.7.4 Properties of natural extension

Suppose $\mathcal{L} \supset \mathcal{D}$, \mathcal{D} avoids sure loss, and \mathcal{E} is its natural extension.

- (a) \mathcal{E} is a proper subset of \mathcal{L}
- (b) \mathcal{E} is coherent relative to \mathcal{L}
- (c) $\mathcal{E} \supset \mathcal{D}$
- (d) $\mathcal{D} = \mathcal{K} \cap \mathcal{E}$ if and only if \mathcal{D} is coherent relative to \mathcal{K}
- (e) \mathcal{E} is the smallest class containing \mathcal{D} that is coherent relative to \mathcal{L} .

3.7.5 Almost-preference

Consider next the representation of beliefs in terms of preferences between gambles. Our model is closely related to the preceding representation in terms of almost-desirable gambles. We will say that gamble X is **almost-preferred** to gamble Y , written as $X \geqslant Y$, just when the gamble $X - Y$ is almost-desirable. The behavioural interpretation is that $X \geqslant Y$ represents a preference for $X + \delta$ over Y , for every positive δ .

To justify this identification of preference with desirability, consider a choice between two compound gambles Z_1 and Z_2 . A fair coin is tossed. If the coin lands ‘heads’ then gamble Z_1 yields $X - Y$ and Z_2 yields zero. If the coin lands ‘tails’ then Z_1 and Z_2 each yield Y . (As usual, all rewards are in probability currency.)

The gamble Z_1 is equivalent to $\frac{1}{2}(X - Y) + \frac{1}{2}Y = \frac{1}{2}X$ in probability currency, whereas Z_2 is equivalent to $\frac{1}{2}Y$. So Z_1 should be preferred to Z_2 if and only if X is preferred to Y . But as Z_1 and Z_2 have the same reward if the coin lands ‘tails’, Your preference between them should be identical to Your preference conditional on ‘heads’. That is, Z_1 should be preferred to Z_2 if and only if $X - Y$ is preferred to zero, which means just that $X - Y$ is desirable. This argument establishes that, when rewards are in probability currency, You should prefer X to Y just when You judge $X - Y$ to be desirable.

Given a relation \geqslant on ordered pairs of gambles, we can define avoiding sure loss and coherence for \geqslant through the corresponding class $\mathcal{D} = \{X - Y : X \geqslant Y\}$ of almost-desirable gambles. So \geqslant avoids sure loss if $\sup \sum_{j=1}^n (X_j - Y_j) \geq 0$ whenever $n \geq 1$ and $X_j \geqslant Y_j$ for $1 \leq j \leq n$.⁶

Say that \geqslant is **coherent relative to \mathcal{J}** (a subset of $\mathcal{L} \times \mathcal{L}$) when \geqslant avoids sure loss and also $X \geqslant Y$ if and only if $(X, Y) \in \mathcal{J}$ and $X - Y \in \mathcal{Z}$. (Here \mathcal{Z} is the natural extension of \mathcal{D} .) This general definition simplifies as follows

when \mathcal{K} is a linear space containing constant gambles and $\mathcal{J} = \mathcal{K} \times \mathcal{K}$, i.e. when You express preferences only between gambles in \mathcal{K} .

3.7.6 Axioms for almost-preference

Let \geqslant be an almost-preference relation on $\mathcal{K} \times \mathcal{K}$, where \mathcal{K} is a linear space containing constant gambles. Then \geqslant is coherent relative to $\mathcal{K} \times \mathcal{K}$ if and only if it satisfies the following six axioms:⁷

- (R0) $\text{not } -1 \geqslant 0$ (avoiding sure loss)
- (R1) if $X, Y \in \mathcal{K}$ and $X \geq Y$ then $X \geqslant Y$ (monotonicity)
- (R2) if $X \geqslant Y$ and $\lambda > 0$ then $\lambda X \geqslant \lambda Y$ (positive homogeneity)
- (R3) if $X \geqslant Y$ and $Y \geqslant Z$ then $X \geqslant Z$ (transitivity)
- (R4) if $X + \delta \geqslant Y$ for all $\delta > 0$ then $X \geqslant Y$ (continuity)
- (R5) $X \geqslant Y$ if and only if $X - Y \geqslant 0$ (cancellation).

Proof. Verify the following consequences of the axioms:

- (a) if $X \geqslant 0$ and $Y \geqslant 0$ then $X + Y \geqslant 0$ (use R3 and R5)
- (b) if $X_j \geqslant Y_j$ for $1 \leq j \leq n$ then $\sum_{j=1}^n X_j \geqslant \sum_{j=1}^n Y_j$ (use (a) and R5)
- (c) if $X \geqslant 0$ then $\sup X \geq 0$ (use R0–R3).

Now suppose $X_j \geqslant Y_j$ for $1 \leq j \leq n$. Then $\sum_{j=1}^n (X_j - Y_j) \geqslant 0$ by (b) and R5, hence \geqslant avoids sure loss by (c).

To show \geqslant is coherent, suppose $X, Y \in \mathcal{K}$ and $X - Y \geq \sum_{j=1}^n \lambda_j (X_j - Y_j)$ where $\lambda_j > 0$ and $X_j \geqslant Y_j$. Then $\sum_{j=1}^n \lambda_j (X_j - Y_j) \geqslant 0$ by R2, R5 and (b), so $X - Y \geqslant 0$ by R1 and R3, and $X \geqslant Y$ by R5. Finally, use R4 to show that if $X, Y \in \mathcal{K}$ and, for every $\delta > 0$, $X - Y + \delta \geq \sum_{j=1}^n \lambda_j (X_j - Y_j)$ for some $\lambda_j > 0$ and $X_j \geqslant Y_j$, then $X + \delta \geqslant Y$ for all $\delta > 0$, so $X \geqslant Y$.

Conversely, R0 is a consequence of avoiding sure loss, R5 holds because $X \geqslant Y$ is equivalent to $X - Y \in \mathcal{E}$, and the other axioms follow from the definition of \mathcal{E} . (For R3, use $X - Z = (X - Y) + (Y - Z)$.) ◆

Axioms R2, R3 and R5 (plus the reflexivity condition $X \geqslant X$ which follows from R1) characterize the **linear partial orderings** on the linear space \mathcal{K} . The coherent relations \geqslant are the linear partial orderings which are non-trivial (satisfy R0), monotone (R1) and continuous (R4).⁸

Given the cancellation axiom R5, which links preference to desirability and was justified in section 3.7.5, the other preference axioms R0–R4 are simple consequences of the desirability axioms D0–D4. The continuity axiom R4 corresponds to the closure axiom D4 and is needed to give a one-to-one correspondence with coherent lower previsions. This axiom is not appropriate for strict preference relations, which we consider next.

3.7 DESIRABILITY AND PREFERENCE

3.7.7 Strict desirability and strict preference

In making decisions we need to know about strict preferences rather than almost-preferences. If You can choose between gambles X and Y , and You strictly prefer X to Y , then You should obviously choose X . But if You almost-prefer X to Y then You may well choose Y ; indeed, You may strictly prefer Y to X !

Again, we will say that X is strictly preferred to Y just when $X - Y$ is strictly desirable. Let \mathcal{D}^+ denote a class of **strictly desirable** gambles. When $X \in \mathcal{D}^+$, You are positively willing to accept X . We could give general definitions of strictly avoiding sure loss, natural extension and coherence for \mathcal{D}^+ , similar to Definitions 3.7.1 and 3.7.2, but instead we will characterize coherence relative to a linear space. To motivate the following axioms, consider which of the almost-desirable gambles in a given class \mathcal{D} must be strictly desirable. Every strictly desirable gamble must allow some possibility of gain (axiom D5 below). The zero gamble is not strictly desirable, but all other non-negative gambles must be (axiom D6). All gambles X in the relative interior of \mathcal{D} , such that $X - \delta \in \mathcal{D}$ for some $\delta > 0$, must be strictly desirable, and then $X - \delta/2$ is also strictly desirable. These are the only gambles in \mathcal{D} that must be strictly desirable. This means that \mathcal{D}^+ (excluding the non-negative gambles) is open under the supremum-norm topology (axiom D7). In fact \mathcal{D}^+ is the union of the non-negative and non-zero gambles with the relative interior of its corresponding class \mathcal{D} , and \mathcal{D} is the relative closure of \mathcal{D}^+ .

3.7.8 Axioms for strict desirability⁹

Let \mathcal{K} be a linear space containing constant gambles and $\mathcal{K} \supset \mathcal{D}^+$. Say that \mathcal{D}^+ is **coherent relative to \mathcal{K}** if it satisfies the five axioms D2, D3, and:

- (D5) if $X \leq 0$ then $X \notin \mathcal{D}^+$ (avoiding partial loss)
- (D6) if $X \in \mathcal{K}$, $X \geq 0$ and $X \neq 0$ then $X \in \mathcal{D}^+$ (accepting partial gains)
- (D7) if $X \in \mathcal{D}^+$ then either $X \geq 0$ or $X - \delta \in \mathcal{D}^+$ for some $\delta > 0$ (openness).

Finally, let $>$ denote a **strict preference** relation between gambles. The interpretation of $X > Y$ is that, when given a choice between X and Y , You are disposed to choose X rather than Y .¹⁰ Strict preference corresponds to strict desirability by $X > Y$ if and only if $X - Y \in \mathcal{D}^+$. That is, You strictly prefer X to Y just when You are positively willing to give up Y in return for X . The following axioms for strict preference correspond to the above axioms for strict desirability.

3.7.9 Axioms for strict preference

Let \mathcal{K} be a linear space containing constant gambles, and $>$ a strict preference relation on $\mathcal{K} \times \mathcal{K}$. Say that $>$ is **coherent relative to** $\mathcal{K} \times \mathcal{K}$ if it satisfies the six axioms R2, R3, R5, and:

- (R6) if $X \geq Y$ then not $Y > X$ (avoiding partial loss)
- (R7) if $X, Y \in \mathcal{K}$, $X \geq Y$ and $X \neq Y$ then $X > Y$ (strict monotonicity)
- (R8) if $X > Y$ then either $X \geq Y$ or $X > Y + \delta$ for some $\delta > 0$ (openness).

3.8 Equivalent models for beliefs

Four new types of models were introduced in the previous section. For each model we have defined a notion of coherence. The next theorem shows that these definitions of coherence are essentially the same as the earlier definition for lower previsions, in that there are natural one-to-one correspondences between the coherent lower previsions and the coherent models for strict and almost-desirability or preference. When coherence is required, the five models therefore convey the same information about beliefs.

3.8.1 Correspondence theorem

Suppose \mathcal{K} is a linear space containing constant gambles.¹ There are one-to-one correspondences between the sets of models of the following types:

1. coherent lower previsions on domain \mathcal{K}
2. classes of almost-desirable gambles \mathcal{D} that are coherent relative to \mathcal{K}
3. almost-preference orderings \geqslant that are coherent relative to $\mathcal{K} \times \mathcal{K}$
4. classes of strictly desirable gambles \mathcal{D}^+ that are coherent relative to \mathcal{K}
5. strict preference orderings $>$ that are coherent relative to $\mathcal{K} \times \mathcal{K}$.

The correspondences involving \underline{P} are defined by²

$$\begin{aligned}\underline{P}(X) &= \max\{\mu: X - \mu \in \mathcal{D}\} = \max\{\mu: X \geqslant \mu\} = \sup\{\mu: X - \mu \in \mathcal{D}^+\} \\ &= \sup\{\mu: X > \mu\}, \\ \mathcal{D} &= \{X \in \mathcal{K}: \underline{P}(X) \geq 0\}, \mathcal{D}^+ = \{X \in \mathcal{K}: X \geq 0 \text{ and } X \neq 0, \text{ or } \underline{P}(X) > 0\}, \\ X \geqslant Y &\text{ when } \underline{P}(X - Y) \geq 0, \\ X > Y &\text{ when } \underline{P}(X - Y) > 0 \text{ or } X \geq Y \text{ and } X \neq Y.\end{aligned}$$

Proof. Verify that these correspondences are bijective and preserve the axioms for coherence given in sections 2.3.3, 3.7.3, 3.7.6, 3.7.8 and 3.7.9. ◆

Two of these correspondences are especially important in the following accounts of decision making and elicitation. We can elicit a coherent lower

prevision \underline{P} through judgements of strict or almost-desirability, defining $\underline{P}(X) = \sup\{\mu: X - \mu \in \mathcal{D}\}$. The elicited \underline{P} can then be used in decision making by examining its implications for strict preferences between gambles, defining $X > Y$ when $\underline{P}(X - Y) > 0$ or when $X \geq Y$ and $X \neq Y$. Note, however, that some information can be lost in going from judgements of desirability, via \underline{P} , to a strict preference ordering. That is because \underline{P} does not contain any information about which ‘marginally desirable’ gambles, those with $\underline{P}(X) = 0$, are really desirable. That issue is discussed in section 3.8.6.

3.8.2 Vacuous previsions

As a simple example of the correspondences, let \underline{P} be the vacuous lower prevision on \mathcal{L} , $\underline{P}(X) = \inf X$. The corresponding models are

- $\mathcal{D} = \{X \in \mathcal{L}: X \geq 0\}$, the class of all non-negative gambles,
- $\mathcal{D}^+ = \{X \in \mathcal{L}: X \geq 0, X \neq 0\}$, the class of non-zero non-negative gambles,
- $X \geqslant Y$ when $X \geq Y$, the natural (pointwise) ordering of \mathcal{L} ,
- $X > Y$ when $X \geq Y$ and $X \neq Y$, the strict pointwise ordering of \mathcal{L} .

The classes \mathcal{D} and \mathcal{D}^+ are the minimal coherent classes of almost- or strictly desirable gambles. Similarly, \geqslant and $>$ are the minimal coherent partial orderings. These facts confirm that the model really is ‘vacuous’.

3.8.3 Completeness

Consider what properties of desirable classes and preference orderings correspond to linearity of previsions. By Theorem 2.8.2, the linear previsions on a linear space \mathcal{K} are just the coherent lower previsions that satisfy the self-conjugacy property $P(X) = -P(-X)$. Using the correspondence in Theorem 3.8.1, these correspond to the classes of almost-desirable gambles \mathcal{D} that are coherent relative to \mathcal{K} (satisfy axioms D0–D4) and also satisfy the **completeness** axiom

- (D8) if $X \in \mathcal{K}$ then $X \in \mathcal{D}$ or $-X \in \mathcal{D}$.

The corresponding completeness axioms for the models \geqslant , \mathcal{D}^+ and $>$ are:

- (R9) if $X, Y \in \mathcal{K}$ then $X \geqslant Y$ or $Y \geqslant X$
- (D9) if $X \in \mathcal{K}$ then $X \in \mathcal{D}^+$ or $-X + \delta \in \mathcal{D}^+$ for all $\delta > 0$
- (R10) if $X, Y \in \mathcal{K}$ then $X > Y$ or $Y + \delta > X$ for all $\delta > 0$.

Assuming coherence, these four completeness axioms are equivalent to the other axioms of precision listed in section 2.8.11.

3.8.4 Corresponding classes of linear previsions

Each of the models \mathcal{D} , \mathcal{D}^+ , \geqslant and $>$ corresponds to a class $\mathcal{M}(P)$ of linear previsions, via the corresponding lower revision P . These classes may be written as

$$\begin{aligned}\mathcal{M}(\mathcal{D}) &= \{P \in \mathcal{P}: \text{for all } X \text{ in } \mathcal{D}, P(X) \geq 0\} \\ \mathcal{M}(\mathcal{D}^+) &= \{P \in \mathcal{P}: \mathcal{P}: \text{for all } X \text{ in } \mathcal{D}^+, X \geq 0 \text{ or } P(X) > 0\} \\ \mathcal{M}(\geqslant) &= \{P \in \mathcal{P}: P(X) \geq P(Y) \text{ whenever } X \geqslant Y\}\end{aligned}$$

and $\mathcal{M}(>) = \{P \in \mathcal{P}: P(X) > P(Y) \text{ whenever } X > Y \text{ and not } X \geq Y\}$.

When \mathcal{D} is coherent (relative to a linear space \mathcal{K}) it is a convex cone, and $\mathcal{M}(\mathcal{D})$ consists of all linear previsions that are positive with respect to this convex cone.³ By correspondence with the natural extension theorem 3.4.1, provided \mathcal{D} avoids sure loss its natural extension is $\mathcal{E} = \{X \in \mathcal{L}: P(X) \geq 0 \text{ for all } P \in \mathcal{M}(\mathcal{D})\}$, and $\mathcal{M}(\mathcal{D}) = \mathcal{M}(\mathcal{E})$. Avoiding sure loss and coherence of \mathcal{D} can be characterized in terms of $\mathcal{M}(\mathcal{D})$ by the following theorem, which corresponds to the lower envelope theorem 3.3.3.

3.8.5 Theorem

Suppose $\mathcal{L} \supset \mathcal{K} \supset \mathcal{D}$, where \mathcal{D} is a class of almost-desirable gambles. The following three conditions are equivalent.

- (a) \mathcal{D} avoids sure loss
- (b) $\mathcal{M}(\mathcal{D})$ is non-empty
- (c) there is a complete class \mathcal{D}' that is coherent relative to \mathcal{L} and contains \mathcal{D} .

The next three conditions are equivalent characterizations of coherence.

- (d) \mathcal{D} is coherent relative to \mathcal{K}
- (e) $\mathcal{M}(\mathcal{D})$ is non-empty and $\mathcal{D} = \{X \in \mathcal{K}: P(X) \geq 0 \text{ for all } P \in \mathcal{M}(\mathcal{D})\}$
- (f) there are classes \mathcal{D}' satisfying (c), and \mathcal{D} is the intersection of \mathcal{K} and all such \mathcal{D}' .

Thus any \mathcal{D} that is coherent relative to \mathcal{K} can be written as an intersection of complete coherent classes, which are just the ‘half-spaces’ of gambles $\{X \in \mathcal{K}: P(X) \geq 0\}$ corresponding to some linear revision P .

Similar results hold for the equivalent models \geqslant , \mathcal{D}^+ and $>$. For example, $\mathcal{M}(\geqslant)$ consists of all linear previsions that are monotone with respect to the partial ordering \geqslant . Also \geqslant avoids sure loss if and only if it can be extended to a complete ordering that is coherent relative to $\mathcal{L} \times \mathcal{L}$, and \geqslant is coherent relative to $\mathcal{K} \times \mathcal{L}$ if and only if it is an intersection of complete coherent orderings of \mathcal{K} .

3.8 EQUIVALENT MODELS FOR BELIEFS

3.8.6 Choice between alternative models

Various mathematical models for beliefs have been defined in this chapter. As well as the five equivalent models listed in Theorem 3.8.1, we have the alternative models:

6. the natural extension \underline{P} of P to \mathcal{L}
7. the natural extension \mathcal{E} of \mathcal{D} (or of \mathcal{D}^+ , \geqslant or $>$)
8. the weak*-compact convex class of linear previsions $\mathcal{M}(P)$
9. the set of extreme points of $\mathcal{M}(P)$.

Any one of these nine models provides exactly the same information about beliefs as any other.⁴ Nevertheless, one kind of model may be more natural or convenient than another for particular purposes.

Models 8 and 9 are the least natural for purposes of interpretation, since all the other models have a direct behavioural interpretation whereas $\mathcal{M}(P)$ is interpreted only through these other models. However, the models $\mathcal{M}(P)$ and $\text{ext } \mathcal{M}(P)$ are the most convenient and tractable for many types of mathematical analysis. It is often useful to regard a coherent lower revision P as the lower envelope of a class of linear previsions, especially when conditional previsions or statistical models are involved. For example, P can be conveniently updated on receiving new evidence by updating the linear previsions in $\mathcal{M}(P)$ or $\text{ext } \mathcal{M}(P)$. The models $\mathcal{M}(P)$ and $\text{ext } \mathcal{M}(P)$ will often be mathematically simpler than the corresponding P , \mathcal{D} or \geqslant , especially when $\text{ext } \mathcal{M}(P)$ is a small finite set.⁵

Because of the close relationship between desirability and preference (3.7.5), the models \mathcal{D} and \geqslant are effectively the same for all practical purposes, as are \mathcal{D}^+ and $>$. The model \mathcal{D} is especially useful in the elicitation of lower previsions because the basic requirement of avoiding sure loss is essentially a condition on \mathcal{D} . In decision problems the most useful model is the strict preference ordering $>$, as outlined in the next section.

We have chosen to present the theory of coherence primarily in terms of a lower revision P . The main reason for this choice is that additive probability and linear expectation are special cases of lower revision. The model P is likely to be more comprehensible than \mathcal{D} or \geqslant to readers familiar with the standard theory of additive probability. Other reasons are that the coherence axioms are simpler for P than for \mathcal{D} or \geqslant ,⁶ and important examples such as inner measure, 0–1 valued probabilities, and the other models in section 2.9, can be defined most simply in terms of P .

There is a strong case, however, for regarding desirability or preference as a more fundamental concept than lower revision. The idea of desirability underlies our interpretation of P and was used in section 2.3 to justify the coherence axioms for P . It will also play a key role in later chapters, especially

in our accounts of elicitation, conglomerability and coherence of statistical models.

Another reason for taking desirability as the fundamental concept is that the class \mathcal{R} of all really desirable gambles is more informative than its induced lower prevision $\underline{P}(X) = \sup\{\mu: X - \mu \in \mathcal{R}\}$. Generally the three classes $\mathcal{D}^+, \mathcal{R}, \mathcal{D}$ will be different, although each induces the same lower prevision \underline{P} . (Assuming coherence, $\mathcal{D} \supset \mathcal{R} \supset \mathcal{D}^+$, but \mathcal{R} need not be open or closed.) The class \mathcal{R} cannot be recovered from \underline{P} because (except for the non-negative gambles) we cannot tell which gambles on the boundary of \mathcal{D} are really desirable. The boundary consists of all gambles X for which $\underline{P}(X) = 0$, and the real number scale for \underline{P} is not rich enough to distinguish between these.⁷

It may seem that the extra information in \mathcal{R} is unimportant in practice because it concerns only ‘infinitesimally fine’ discriminations between marginally desirable gambles. This extra information is essential, however, for defining previsions conditional on events of upper probability zero, about which the unconditional lower prevision \underline{P} is completely uninformative. Conditional previsions cannot always be recovered from \underline{P} , but they can always be defined in terms of \mathcal{R} , as described in Appendix F.⁸ The whole theory, and especially the theory of statistical inference, seems simplest in terms of the model \mathcal{R} .

3.9 Decision making

Here we outline our approach to decision making, which is based on the strict preference ordering $>$ discussed in the previous sections.¹ The preference ordering is constructed from separate assessments of previsions and utilities. In general, both previsions and utilities may be imprecise, and we construct a partial preference ordering. Here we will concentrate on the special case where the utility function associated with each possible action is precise (or linear), in which case it can be identified with a gamble.² The last example of the section indicates that the general case of imprecise utilities is somewhat more complicated.

3.9.1 Correspondence between actions and gambles

Let \mathbb{A} denote a set of feasible actions from which a choice is to be made. To help assess preferences between the actions, You introduce a possibility space Ω that is relevant to the outcomes of the actions.³ A reward $c(\omega, a)$ is associated with each pair $\omega \in \Omega$ and $a \in \mathbb{A}$; this describes the consequences of taking action a when ω is the true state.³ We assume that the value for You of these consequences can be represented by a bounded linear utility function U . Then each action $a \in \mathbb{A}$ can be evaluated through the gamble

3.9 DECISION MAKING

$X_a \in \mathcal{L}(\Omega)$ defined by $X_a(\omega) = U(c(\omega, a))$. You will prefer one action b to another action a just when You prefer the gamble X_b to the gamble X_a .

Suppose Your beliefs about Ω are represented by a coherent lower prevision $(\Omega, \mathcal{L}, \underline{P})$. (If necessary, extend \underline{P} to \mathcal{L} by natural extension.) Then \underline{P} tells us something about preferences between gambles, through the strict partial preference ordering $>$ that corresponds to \underline{P} . By Theorem 3.8.1, that is defined by: $X_b > X_a$ if and only if $\underline{P}(X_b - X_a) > 0$, or $X_b \geq X_a$ and $X_b \neq X_a$. Thus \underline{P} induces a partial ordering on the set of feasible actions \mathbb{A} .

Because actions a are evaluated only through the gambles X_a , we can simplify the notation by regarding the choice of an action from \mathbb{A} as the choice of a gamble from $\mathcal{K} = \{X_a: a \in \mathbb{A}\}$.⁴ If $X_b > X_a$ then gamble X_a (and its corresponding action a) can be eliminated from consideration; X_b is strictly better. So the only gambles that are worth considering are those that are maximal under the partial ordering $>$. In terms of \underline{P} , the maximal gambles are defined as follows.

3.9.2 Definitions

Suppose \underline{P} is a coherent lower prevision on \mathcal{L} , \bar{P} is its conjugate upper prevision and $\mathcal{L} \supset \mathcal{K}$. Let $X \in \mathcal{K}$. Say that X is **inadmissible** in \mathcal{K} when there is $Y \in \mathcal{K}$ such that $Y \geq X$ and $Y \neq X$. Otherwise X is **admissible** in \mathcal{K} . Say that X is **maximal** in \mathcal{K} under \underline{P} when X is admissible in \mathcal{K} and $\underline{P}(X - Y) \geq 0$ for all $Y \in \mathcal{K}$.

When X is not maximal in \mathcal{K} , either it is inadmissible or there is $Y \in \mathcal{K}$ such that $\underline{P}(Y - X) > 0$, so that $Y > X$. Thus the maximal gambles under \underline{P} are just the gambles that are maximal (undominated) under the partial ordering $>$ induced by \underline{P} . For an admissible gamble X to be maximal in \mathcal{K} it is sufficient, but not necessary, that either $\underline{P}(X) \geq \underline{P}(Y)$ or $\bar{P}(X) \geq \bar{P}(Y)$ for every $Y \in \mathcal{K}$. It is necessary, but not sufficient, that $\bar{P}(X) \geq \bar{P}(Y)$ for all $Y \in \mathcal{K}$.

The first technical question concerns the existence of maximal gambles. There can be no gambles which are maximal, or even admissible, when \mathcal{K} is open. It can be shown that if \mathcal{K} is compact under the supremum-norm topology, and \underline{P} is coherent, then there are maximal gambles in \mathcal{K} under \underline{P} .⁵ (This applies if \mathcal{K} is finite.)

3.9.3 Indeterminacy

You can always reduce the set of options \mathcal{K} to the subset of admissible gambles, which are just the gambles that are maximal under the vacuous lower prevision. When \underline{P} is non-vacuous You can further reduce the set of options to the subset of gambles that are maximal under \underline{P} . (We assume that these subsets are non-empty.) Typically \underline{P} is imprecise, $>$ is not a

complete ordering of \mathcal{K} and there is more than one maximal gamble in \mathcal{K} . In that case, since \underline{P} does not determine any preferences between maximal gambles, You cannot identify (through \underline{P}) a unique optimal gamble. You may sometimes be able to discriminate between maximal gambles on other grounds (see section 5.6.6), but often You will be simply unable to say that one maximal gamble is ‘better than’ another. Of course, that will not prevent You from choosing a single maximal gamble, arbitrarily if necessary.

In the extreme case where there is no information concerning the probabilities of states ω , \underline{P} is vacuous, all admissible gambles are maximal under \underline{P} , and there seems to be no basis for preferring one admissible gamble to another. The failure to determine a uniquely optimal gamble simply reflects the absence of information about Ω which could be used to discriminate between gambles. More generally, the incompleteness of the preference ordering reflects the imprecision in \underline{P} .⁶

3.9.4 Bayes gambles

When P is a linear prevision, the maximal gambles under P are the admissible gambles X that satisfy $P(X) \geq P(Y)$ for all $Y \in \mathcal{K}$. Any gamble which maximizes $P(Y)$ over $Y \in \mathcal{K}$ is called a **Bayes gamble** under P , as it is ‘optimal’ in the standard Bayesian sense of maximizing expected utility. (The corresponding action is called a **Bayes action**.) Thus the maximal gambles under P are just the admissible Bayes gambles. If there is a unique Bayes gamble under P then it must be admissible and it is therefore uniquely maximal under P . In general, even when beliefs are represented by a linear prevision P , there may be many maximal gambles and many Bayes gambles.⁷

An admissible gamble X is maximal under a coherent lower prevision \underline{P} just when, for each $Y \in \mathcal{K}$, there is $P \in \mathcal{M}(\underline{P})$ such that $P(X) \geq P(Y)$. If X is maximal under some $P \in \mathcal{M}(\underline{P})$ then $P(X) \geq P(Y)$ for all $Y \in \mathcal{K}$, so X is maximal under \underline{P} . We show next that the converse holds (all gambles maximal under \underline{P} are maximal under some $P \in \mathcal{M}(\underline{P})$), provided the set of options \mathcal{K} is convex. That can always be achieved by **randomization**. That is, given a basic set of options \mathcal{K}_0 , You can introduce a randomizing device that selects one gamble from a finite subset of \mathcal{K}_0 according to a known probability distribution. If all such randomizing distributions can be chosen, You can effectively choose any gamble in the convex hull of \mathcal{K}_0 .⁸ So it seems reasonable to assume that the set of options \mathcal{K} is convex.

3.9.5 Maximality theorem

Suppose that \underline{P} is a coherent lower prevision defined on \mathcal{L} , \bar{P} is its conjugate upper prevision, and \mathcal{K} is a convex subset of \mathcal{L} . Then $\bar{P}(X - Y) \geq 0$ for all

3.9 DECISION MAKING

$Y \in \mathcal{K}$ if and only if there is $P \in \mathcal{M}(\underline{P})$ such that $P(X - Y) \geq 0$ for all $Y \in \mathcal{K}$, i.e. X is a Bayes gamble under P . Hence, a gamble is maximal in \mathcal{K} under \underline{P} if and only if it is maximal in \mathcal{K} under some P in $\mathcal{M}(\underline{P})$.

*Proof.*⁹ Let $\mathcal{W} = \{X - Y: Y \in \mathcal{K}\}$. Since \mathcal{K} is convex, so is \mathcal{W} . It suffices to prove that $\bar{P}(Z) \geq 0$ for all $Z \in \mathcal{W}$ if and only if there is $P \in \mathcal{M}(\underline{P})$ such that $P(Z) \geq 0$ for all $Z \in \mathcal{W}$. The second condition is sufficient since $\bar{P}(Z) \geq P(Z)$ when $P \in \mathcal{M}(\underline{P})$.

To prove necessity, let $\mathcal{D} = \mathcal{W} \cup \{V \in \mathcal{L}: P(V) \geq 0\}$. Suppose $Z_i \in \mathcal{W}$ and $P(Z_i) \geq 0$. Let $W = m^{-1} \sum_{i=1}^m Z_i$, so $W \in \mathcal{W}$ and $\bar{P}(W) \geq 0$. Hence $\sup(\sum_{i=1}^m Z_i + \sum_{j=1}^n V_j) \geq \bar{P}(mW + \sum_{j=1}^n V_j) \geq \bar{P}(mW) + \sum_{j=1}^n \bar{P}(V_j) \geq m\bar{P}(W) \geq 0$. Thus this proves that the class \mathcal{D} satisfies condition (a) of Lemma 3.3.2, and so there is a linear prevision P on \mathcal{L} such that $P(U) \geq 0$ for all $U \in \mathcal{D}$. For all $Z \in \mathcal{L}$, $Z - P(Z) \in \mathcal{D}$ giving $P(Z) \geq P(Z)$. Thus $P \in \mathcal{M}(\underline{P})$. Since $\mathcal{D} \supset \mathcal{W}$, $P(Z) \geq 0$ for all $Z \in \mathcal{W}$. This establishes the result. ◆

When \underline{P} is the vacuous lower prevision, maximality under \underline{P} is equivalent to admissibility and the maximality theorem implies the following.

3.9.6 Corollary

If \mathcal{K} is convex then every gamble that is admissible in \mathcal{K} is a Bayes gamble. (Indeed, if X is not a Bayes gamble then it is uniformly inadmissible, in the sense that $X \leq Y - \delta$ for some $Y \in \mathcal{K}$ and $\delta > 0$.)

The maximality theorem shows that, when the set of optional gambles is convex, the maximal gambles under \underline{P} are just the admissible gambles that are Bayes gambles under some dominating linear prevision. In that case, all reasonable actions are Bayes actions. One way to choose a maximal gamble (or action) is to first choose a unique P from $\mathcal{M}(\underline{P})$ and then choose a gamble that is maximal under P . The theorem shows that all maximal gambles could be obtained in this way. Typically, however, it will be more difficult to choose a unique linear prevision than to choose a unique maximal gamble, because the same maximal gamble will be Bayes under many different linear previsions.

3.9.7 Minimax gambles

Another way of choosing a maximal gamble from \mathcal{K} is to select an admissible gamble which maximizes $\underline{P}(X)$ over $X \in \mathcal{K}$. Any such gamble X_0 is called an admissible \underline{P} -minimax gamble.¹⁰ Assuming \underline{P} is coherent, X_0 must be maximal since $\bar{P}(X_0 - Y) \geq \underline{P}(X_0) - \underline{P}(Y) \geq 0$ for all $Y \in \mathcal{K}$. Provided the set of options \mathcal{K} is convex and compact, it must contain admissible

\underline{P} -minimax gambles. In that case, by the minimax theorem (Appendix E6), there are $X_0 \in \mathcal{K}$ and $P_0 \in \mathcal{M}(\underline{P})$ such that X_0 is an admissible \underline{P} -minimax gamble, X_0 is a Bayes gamble under P_0 , $P_0(X_0) = \underline{P}(X_0)$, and P_0 is ‘least-favourable’ in $\mathcal{M}(\underline{P})$ in the sense that it minimizes $\max\{\underline{P}(X) : X \in \mathcal{K}\}$ over all $P \in \mathcal{M}(\underline{P})$.

This definition of \underline{P} -minimax gambles generalizes the well-known definition of minimaxity, which is recovered when \underline{P} is taken to be vacuous. At the other extreme, when P is a linear prevision, the P -minimax gambles are just the Bayes gambles under P .

The \underline{P} -minimax gambles may be appealing because they have the highest buying prices $\underline{P}(X)$ of the gambles in \mathcal{K} . However, there does not seem to be any good reason to prefer the \underline{P} -minimax gambles to the other maximal gambles. Experience with minimax rules in statistical decision problems suggests that \underline{P} -minimax gambles can be difficult to find and are not particularly good choices in general.¹¹ The admissible \underline{P} -minimax gambles are at least ‘reasonable’ choices in the sense that no other gamble is strictly preferred to them under \underline{P} .

3.9.8 Statistical decision problems

As a special case of the general decision problem, suppose that the state ω is a statistical parameter, the possible decisions $d \in \mathbb{A}$ are actually **decision rules** which specify an action $d(x)$ for each possible outcome x of a statistical experiment, and the gamble X_d specifies the expected utility from using rule d conditional on the possible values of ω . (Here $-X_d$ is usually called the **risk function** of the decision rule d .)

The lower prevision \underline{P} now represents prior beliefs about the parameter ω , before the outcome x is observed. When \underline{P} is vacuous, the maximal decision rules are just the admissible rules, and our definition of admissibility agrees with the standard definition in terms of dominance of risk functions. By Corollary 3.9.6, all admissible rules (in a class that is closed under randomization) are Bayes decision rules.¹² In **classical decision theory**, attention is restricted to the class of admissible rules, usually a large class, but no further preferences are specified amongst the admissible rules.¹³ Classical decision theory can therefore be regarded as a special case of our approach in which the prior \underline{P} is vacuous.

When the prior P is linear, the maximal decision rules are just the admissible Bayes decision rules under P , those which maximize the prior expected utility $P(X_d)$ or minimize the prior expected risk. In **Bayesian decision theory**, attention is restricted to the class of Bayes decision rules under a single linear prevision P . This class is usually much smaller than the class of all admissible rules.

3.9 DECISION MAKING

Our approach includes both classical and Bayesian decision theories as extreme cases. Both theories seem inadequate for most practical problems. A vacuous prior will usually be unrealistically imprecise, and a linear prior will usually be unrealistically precise. You will usually have some prior information about ω that can be used to assess a coherent lower prevision \underline{P} and form some preferences between admissible rules, but not enough information to determine a complete preference ordering.

The maximality theorem 3.9.5 relies on two assumptions: that the set of options \mathcal{K} is convex, and that the utility functions associated with actions are precise. To end our discussion of decision making we give two examples to show that, when either assumption is violated, there can be reasonable actions that are not Bayes actions.¹⁴

3.9.9 Example

Let $\Omega = \{\omega_1, \omega_2\}$ and $\mathcal{K} = \{X_1, X_2, X_3\}$, where the gambles $X = (X(\omega_1), X(\omega_2))$ are $X_1 = (1, 1)$, $X_2 = (3, 0)$ and $X_3 = (0, 3)$. Let \underline{P} be vacuous. Each gamble is admissible in \mathcal{K} , so X_1 is a maximal gamble in \mathcal{K} under \underline{P} . But X_1 is not maximal under any linear prevision P , since X_2 is better when $P(\{\omega_1\}) > \frac{1}{3}$ and X_3 is better when $P(\{\omega_1\}) < \frac{2}{3}$. (Of course X_1 cannot be maximal under \underline{P} in the convex hull of \mathcal{K} . Gambles such as $\frac{1}{2}X_2 + \frac{1}{2}X_3$ are preferred to X_1 .)

3.9.10 Example of imprecise utilities

Again let $\Omega = \{\omega_1, \omega_2\}$ with \underline{P} vacuous. Consider three actions a_1, a_2, a_3 , where the precise utility functions associated with a_1 and a_2 are the gambles X_1 and X_2 in the previous example. We can replace X_3 by an imprecise utility function in such a way that a_1 is still not a Bayes action but it is no longer dominated by any randomized action. The imprecise utilities associated with a_3 will be represented by the class of linear utility functions of the form $X(\omega_1) = 1 - \mu$ and $X(\omega_2) = 1 + \mu$, where μ takes all values in the interval $[0.2, 0.5]$.

Each choice of precise probability and precise utility function, specified by $\rho = P(\{\omega_1\})$ and by μ , determines precise expected utilities 1 , 3ρ and $1 + \mu(1 - 2\rho)$ for the actions a_1, a_2 and a_3 . Hence, the values ρ and μ determine a complete preference ordering on the class \mathbb{A} of all randomized actions. The partial preference ordering on \mathbb{A} that is determined by the imprecise probabilities and imprecise utilities is thus taken to be the intersection of these complete orderings, over all pairs (ρ, μ) with $0 \leq \rho \leq 1$ and $0.2 \leq \mu \leq 0.5$. This definition can be justified through a generalized

notion of natural extension.¹⁵ We find that:

1. Action a_1 is not a Bayes action for any $0 \leq \rho \leq 1$ and $0.2 \leq \mu \leq 0.5$, since a_2 is better than a_1 when $\rho > \frac{1}{3}$, and a_3 is better than a_1 when $\rho < \frac{1}{2}$ and $\mu > 0$.
2. a_1 is **minimax** in **A**, in the sense that it maximizes the lower expected utility, defined as the minimum expected utility over pairs (ρ, μ) . The lower expected utility of a_1 is 1. Let b_λ denote a randomized choice of a_2 or a_3 , with probabilities λ and $1 - \lambda$. Given (ρ, μ) , b_λ has expected utility $3\rho\lambda + (1 + \mu(1 - 2\rho))(1 - \lambda)$. This takes the values $1.2(1 - \lambda)$ when $\rho = 0$ and $\mu = 0.2$, and $0.5 + 2.5\lambda$ when $\rho = 1$ and $\mu = 0.5$. Hence the lower expected utility of b_λ is less than 1 for all λ . It follows that a_1 is uniquely minimax in **A**.
3. a_1 is maximal in **A** under the partial preference ordering, i.e. no randomized action is preferred to a_1 . This follows from (2): a_1 is preferred to any other action a under some pair (ρ, μ) , so a cannot be preferred to a_1 in the partial ordering.

Here the action a_1 seems reasonable, even though it is not optimal under any combination of precise probabilities and utilities, because there is no other (possibly randomized) action that is preferred to it. When both probabilities and utilities are imprecise, as in many practical decision problems, it can be reasonable to choose actions that are not Bayes actions.

CHAPTER 4

Assessment and elicitation

There are many ways of assessing imprecise probabilities. Various methods of assessment and elicitation are described in this chapter. All these methods can be regarded as special cases of a general elicitation procedure, which accepts a wide variety of probability judgements and constructs an imprecise probability model by natural extension.

It is useful to distinguish between **assessment** and **elicitation** of probabilities. Elicitation is the process by which beliefs (pre-existing behavioural dispositions) are measured, through explicit judgements and choices. Assessment is the process by which probabilities are constructed from the available evidence. Both processes result in probability models, but the aim in elicitation is to model pre-existing beliefs, whereas the aim in assessment is to formulate rational beliefs. Assessment is more fundamental than elicitation for several reasons: it is important for Your beliefs and probabilities to properly reflect Your evidence; and You will not always have non-vacuous beliefs waiting to be elicited.

Assessment and elicitation are essential topics in a theory of probability, especially one that aims to be applicable. The methods described in this chapter are useful in both assessment and elicitation. In both tasks You must make some probability judgements, whether these express existing beliefs or are formed by analysing the evidence.

Here is an outline of the chapter. A general procedure for eliciting unconditional probabilities is described in section 4.1. This admits any probabilistic judgements that can be regarded as judgements of desirability. Provided the judgements avoid sure loss, they generate an imprecise probability model by natural extension. This procedure can be compared with Bayesian elicitation procedures, and with the operational measurement procedures described in Appendix H. The basic problems, for the general procedure and the special types of judgement considered in later sections, are to (a) explain the behavioural meaning of the judgements, (b) determine if they avoid sure loss, and (c) compute their natural extension. In section 4.2 we show how to solve problems (b) and (c) when both Ω and the set of judgements are finite.

A football example is used throughout the chapter to illustrate various kinds of assessments and models. Because Ω has only three states, imprecise probability models can be displayed graphically, in the probability simplex described in section 4.2. This gives some geometrical insight into the effects of different judgements, and it also illustrates the general theory of Chapters 2 and 3. (The example is developed in Appendix I.)

In section 4.3, assessment is viewed as a sequential process. We describe some basic steps which may be taken to modify the current probability model. These steps include adding or removing judgements of desirability, combining assessments from different sources, and modifying the possibility space in various ways such as refinement, multivalued mapping or conditioning. Complex, highly flexible assessment strategies can be built up from these simple steps.

The last three sections (4.4 to 4.6) describe some special types of probability judgements. The most common judgements in everyday life are classificatory, that an event is probable, or comparative, that one event is at least as probable as another. These judgements are easy to understand and elicit. Classifications of the probable events (section 4.4) generate highly imprecise models, except when Ω is large. Comparative probability orderings (section 4.5) can produce much greater precision, especially when the orderings are complete, or when events in Ω are compared with standard events whose probabilities are precise.

Some other types of assessment are described in section 4.6, including imprecise assessments of probabilities, probability ratios, density functions, distribution functions and quantiles. Some general types of models (intervals of measures, neighbourhoods of additive probabilities, and classes of conjugate priors) are also described. These may be especially useful for modelling prior beliefs about a statistical parameter.

The assessment methods described in this chapter are by no means sufficient to cope with all practical problems. Assessments of conditional probabilities, and structural judgements such as independence, are needed in most problems. These are discussed in later chapters, especially Chapters 6 and 9.

4.1 A general elicitation procedure

First we describe a general procedure for eliciting beliefs.¹ There are three steps in the procedure, which involve (a) eliciting any probability judgements that express (or can be interpreted as expressing) the desirability of specific gambles, (b) checking that Your judgements avoid sure loss, (c) constructing a coherent lower prevision, defined on all gambles, by natural extension of

the class of desirable gambles. The required concepts (desirability, avoiding sure loss and natural extension) have been discussed in the two previous chapters. Here we will explain their role in elicitation, and discuss some issues of interpretation and practical implementation.

4.1.1 Direct judgements

The general elicitation procedure admits a wide variety of probability judgements, allowing You to express Your beliefs in whatever forms are most natural and meaningful to You. You may make, for example, classificatory or comparative probability judgements, that an event A is probable or that A is at least as probable as B . Indeed, we permit any judgements that can be interpreted as statements that specific gambles are almost desirable for You. Such judgements will be called **direct judgements** of desirability.

For example, ' A is probable' can be interpreted as an expression of willingness to accept any bet on A at odds of better than even money, or any gamble $A - \mu$ where $\mu < \frac{1}{2}$, so that the gamble $A - \frac{1}{2}$ is almost desirable. In this way, the first step in elicitation produces a class \mathcal{D} of gambles that You judge to be almost desirable. Because of the correspondence between desirability, preference and prevision (section 3.8), Your judgements can be equivalently interpreted as statements about Your preferences or lower previsions.

To illustrate the wide range of judgements admitted in elicitation, we list some of the simplest types of direct judgement, together with their translations into statements about the class \mathcal{D} of almost-desirable gambles and the corresponding lower prevision \underline{P} . In the list A, B, C are events, X, Y are gambles, and λ is a specified real number.

1. A is probable $\rightarrow A - \frac{1}{2} \in \mathcal{D}, \underline{P}(A) \geq \frac{1}{2}$
2. A is at least as probable as $B \rightarrow A - B \in \mathcal{D}, \underline{P}(A - B) \geq 0$
3. indifference between A and $B \rightarrow A - B \in \mathcal{D}$ and $B - A \in \mathcal{D}, \underline{P}(A - B) = \bar{P}(A - B) = 0$
4. A is at least λ times as probable as $B \rightarrow A - \lambda B \in \mathcal{D}, \underline{P}(A - \lambda B) \geq 0$
5. A is probable conditional on $B \rightarrow B(A - \frac{1}{2}) \in \mathcal{D}, \underline{P}(B(A - \frac{1}{2})) \geq 0$
6. conditional on C , A is at least as probable as $B \rightarrow C(A - B) \in \mathcal{D}, \underline{P}(C(A - B)) \geq 0$
7. the odds against A are at least λ to 1 $\rightarrow A^c - \lambda A \in \mathcal{D}, \bar{P}(A) \leq (1 + \lambda)^{-1}$
8. X is (almost) desirable $\rightarrow X \in \mathcal{D}, \underline{P}(X) \geq 0$
9. willingness to pay price λ for $X \rightarrow X - \lambda \in \mathcal{D}, \underline{P}(X) \geq \lambda$
10. willingness to sell X for $\lambda \rightarrow \lambda - X \in \mathcal{D}, \bar{P}(X) \leq \lambda$
11. X is at least as good as $Y \rightarrow X - Y \in \mathcal{D}, \underline{P}(X - Y) \geq 0$

12. X is desirable conditional on $B \rightarrow BX \in \mathcal{D}$, $\underline{P}(BX) \geq 0$
13. willingness to pay λ for X conditional on $B \rightarrow B(X - \lambda) \in \mathcal{D}$, $\underline{P}(B(X - \lambda)) \geq 0$.

More complicated judgements can be constructed from these simple judgements,² e.g.:

14. the lower variance of X is at least $\lambda \rightarrow (X - \mu)^2 - \lambda \in \mathcal{D}$ and $\underline{P}((X - \mu)^2) \geq \lambda$ for every real μ .

It is important to check that these behavioural translations do capture the intended meaning of Your probability judgements. This may require some explanation of probability currency or other utility scale, and of the notion of a ‘gamble’.

The general elicitation procedure allows any combination of different types of direct judgements. You are free to make as many or as few judgements as You wish, as long as Your judgements reflect genuine beliefs. It is not necessary to choose between two gambles X and Y , for instance, when You have no definite preference between them.

The general procedure is highly flexible, but it will usually be necessary, especially in more complex problems, to structure the elicitation process in some way, by restricting attention to special types of probability judgements, assessment strategies or models. Some structuring may be necessary because (a) inexperienced assessors can make only simple kinds of probability judgements (sections 4.4, 4.5); (b) it can be difficult to verify that an unstructured set of judgements avoids sure loss and to compute its natural extension (section 4.2); (c) the resulting probability model may be intractable (section 4.6.8); (d) the possibility space Ω may have high dimension (section 4.6.10); or (e) some types of judgement may be more useful than others in reducing indeterminacy or indecision (section 4.3).

4.1.2 Other types of judgement

Many of the probability judgements encountered in practice can be regarded as direct judgements, that particular gambles are desirable. There are, however, some important kinds of probability judgement that are not direct judgements, but which might be admitted in elicitation.³

1. Some **conditional** probability judgements cannot be fully interpreted in terms of an unconditional lower prevision. In particular, the translation of judgement 13 of section 4.1.1 is inadequate when the conditioning event B has upper probability zero. Such judgements can be adequately translated using conditional lower previsions, discussed in Chapter 6.
2. As noted in section 3.8.6, judgements of **real desirability**, rather than

4.1 A GENERAL ELICITATION PROCEDURE

- almost desirability, cannot be fully interpreted in terms of a lower prevision.⁴
3. Various **structural** judgements, notably of independence, permutability and exchangeability, have an important role in most practical problems. Whereas direct judgements can be written in the form ' $X \in \mathcal{E}$ ', structural judgements can be written as constraints 'if $X \in \mathcal{E}$ then $Y \in \mathcal{E}$ ' on the model \mathcal{E} . Our elicitation procedure is extended in section 9.6 to admit structural judgements.
 4. It may be possible to directly assess the **degree of imprecision** $\Delta(X) = \bar{P}(X) - \underline{P}(X)$, especially when this depends in a simple way on the amount of available information or the degree of conflict between several sources of information.⁵
 5. Judgements of lower bounds for **variances** (14) can be regarded as direct judgements, but judgements of upper bounds cannot.⁶
 6. Bayesian sensitivity analysts elicit a class of probability measures \mathcal{M} by imposing constraints, such as symmetry, unimodality, or specified quantiles, on the individual measures in \mathcal{M} . Such constraints can often be given a behavioural interpretation, by translating them into direct judgements of desirability. It is essential, in evaluating whether the constraints are reasonable, to consider their behavioural meaning.⁷

4.1.3 Natural extension

The first stage of the elicitation procedure produces a class \mathcal{D} of gambles that You judge almost desirable. The second step is to check that \mathcal{D} avoids sure loss. Otherwise, there is a finite combination of almost-desirable gambles from which You must lose more than some positive amount. In that case You must reconsider the judgements made during elicitation. Did all Your judgements express genuine preferences? Do You understand and accept the behavioural translations of Your judgements? Are You willing to act on these judgements, knowing that some combination of such actions produces a sure loss?

Provided \mathcal{D} avoids sure loss, no further input is needed from You. The final step is to compute the natural extension \mathcal{E} (Definition 3.7.2) and its corresponding lower prevision \underline{P} , which is defined for all gambles. The natural extensions \mathcal{E} and \underline{P} are coherent, while \mathcal{D} need not be. Indeed, there is no reason to expect coherence of \mathcal{D} since You are merely asked in elicitation to make some probability judgements, without following through all the implications of these judgements, e.g. that finite sums of desirable gambles are desirable.

The natural extension of \mathcal{D} can be most simply described, using the notation of section 3.8.4, in terms of the class $\mathcal{M} = \mathcal{M}(\mathcal{D})$, which consists

of all linear previsions P such that $P(X) \geq 0$ for all X in \mathcal{D} . Provided \mathcal{D} avoids sure loss, \mathcal{M} is non-empty and the natural extension \mathcal{E} consists of all gambles X such that $P(X) \geq 0$ for all P in \mathcal{M} . The corresponding lower prevision \underline{P} is defined by $\underline{P}(X) = \min\{P(X): P \in \mathcal{M}\}$.

Each gamble X in \mathcal{D} determines a closed convex half-space \mathcal{M}_X , consisting of those linear previsions P for which $P(X) \geq 0$. Thus \mathcal{M} is the intersection of half-spaces \mathcal{M}_X , over all X in \mathcal{D} .⁸

4.1.4 Football example

Consider a football match in which the three possible outcomes are win (W), draw (D) and loss (L) for the home team. To simplify the notation we use W to stand for the corresponding event (subset of Ω) and for all the gamble that pays 1 unit if ‘win’ occurs, as well as for the possible outcome (element of Ω). Suppose that Your beliefs about the match are expressed through the following simple probability judgements, and translated into desirability of the gambles X_1, X_2, X_3, X_4 .

1. win is improbable $\rightarrow X_1 = \frac{1}{2} - W = D + L - \frac{1}{2}$
2. win is at least as probable as draw $\rightarrow X_2 = W - D$
3. draw is at least as probable as loss $\rightarrow X_3 = D - L$
4. the odds against loss are no more than 4 to 1 $\rightarrow X_4 = L - 0.2$.

These kinds of imprecise judgements might be reasonable if You had some knowledge about the outcomes of previous matches played under similar conditions. Since no two matches are played under identical conditions, it might be difficult to make more precise assessments.

If only these four judgements are made, the formal model is $\mathcal{D} = \{X_1, X_2, X_3, X_4\}$. This avoids sure loss since it is consistent with the uniform distribution. The class of linear previsions $\mathcal{M}(\mathcal{D})$ is the intersection of the four half-spaces

$$\begin{aligned}\mathcal{M}_1 &= \{P \in \mathcal{P}: P(W) \leq 0.5\}, & \mathcal{M}_2 &= \{P: P(W) \geq P(D)\}, \\ \mathcal{M}_3 &= \{P: P(D) \geq P(L)\} & \text{and} & \mathcal{M}_4 = \{P: P(L) \geq 0.2\}.\end{aligned}$$

The extreme points of $\mathcal{M}(\mathcal{D})$ are found in section 4.2.2. They correspond to the four probability mass functions

$$\begin{aligned}P_1 &= (P_1(W), P_1(D), P_1(L)) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), & P_2 &= (0.5, 0.25, 0.25), \\ P_3 &= (0.5, 0.3, 0.2) & \text{and} & P_4 = (0.4, 0.4, 0.2).\end{aligned}$$

The natural extension \underline{P} is just the lower envelope of these four linear previsions. It can be partially summarized through the lower and upper

probabilities (\underline{P}, \bar{P}) of the possible outcomes, which are $(0.33, 0.5)$ for win, $(0.25, 0.4)$ for draw, $(0.2, 0.33)$ for loss.⁹

Because Ω has only three elements, the class of linear previsions $\mathcal{M}(\mathcal{D})$ can be represented graphically as a subset of the two-dimensional probability simplex. This graphical representation gives some insight into the three steps of the elicitation procedure. It is presented in section 4.2, together with the computation of the natural extension.

4.1.5 Interpretations of the elicited model

The meaning of the elicited models \mathcal{E} , \underline{P} and \mathcal{M} is clear. They represent what You have revealed about Your beliefs through Your explicit judgements.¹⁰ The imprecision of these models reflects both indeterminacy in Your beliefs and incompleteness of the models (in the sense of section 2.10.3); Your explicit judgements may not fully reveal Your beliefs. This kind of incompleteness seems unavoidable in practice.

Because of their incompleteness, our models can be given a sensitivity analysis interpretation, by assuming that You have determinate beliefs which could (ideally) be modelled by some linear prevision in $\mathcal{M}(\mathcal{D})$. But we will see in Chapter 9 that, when the elicitation procedure is extended to allow structural judgements, it may be possible to model these judgements by a coherent lower prevision \underline{P} but not by any linear prevision. In such cases a sensitivity analysis interpretation of \underline{P} is ruled out.

There is a temptation, which we have resisted, to construe elicitation as an operational measurement procedure, by exchanging some of the gambles that You judge to be desirable. Several operational measurement schemes have been advocated by Bayesians,¹¹ and others are outlined in Appendix H. Although these schemes may be useful in special contexts, they seem impracticable in general. It is useful in elicitation to consider the behavioural meaning of probability judgements, that You would be willing (in hypothetical circumstances) to accept certain gambles, but it is unnecessary (and usually impracticable) to actually exchange the gambles.¹²

4.1.6 Bayesian elicitation

Bayesian methods of elicitation can be regarded as special cases of the general elicitation procedure, in which restrictions are imposed to determine a complete class \mathcal{E} or a unique linear prevision.¹³ In Example 4.1.4, for instance, You might be asked to assess two-sided betting rates (additive probabilities) for the events W, D, L . Even in such a simple problem it may be difficult to make precise assessments, except by choosing the numbers in a somewhat arbitrary way.

It is difficult also to determine a unique linear prevision P by the special methods described later in this chapter. For example, an elicited comparative probability ordering of finitely many events will typically not determine a unique P , even when the ordering is complete. It is necessary to extend the ordering to infinitely many events by introducing ‘extraneous measurement’ events (section 4.5.7). The assessment task is then much more difficult as it requires infinitely many comparisons of arbitrary precision. If You try to determine a unique distribution function for a gamble by assessing its quantiles (section 4.6.7), You must either assess infinitely many quantiles or else (as is common in practice) interpolate in an arbitrary way between the elicited quantiles.

In general, Bayesians cannot make the precise assessments needed to determine a unique P without some kind of arbitrary choice. On the other hand, Bayesians must be careful not to over-determine the linear prevision. When they make more judgements than are required to determine P (by using several elicitation methods, for example), these will typically incur sure loss.¹⁴

It is not surprising that, in practice, attempts to elicit precise probabilities often produce judgements that are unstable or incoherent. The Bayesian approach does not distinguish judgements that reflect genuine preferences and discriminations from others that are arbitrary. In contrast, the general elicitation procedure allows You to abstain from arbitrary judgement.

4.2 Finitely generated models and simplex representations

4.2.1 Finitely generated models

In this section we discuss the computational aspects of the general elicitation procedure, for the case in which Ω is effectively finite. As in section 4.1, let \mathcal{D} be the class of gambles judged almost desirable, and suppose that \mathcal{D} avoids sure loss so that it has coherent natural extension \mathcal{E} with corresponding models \underline{P} and \mathcal{M} . Say that the model \mathcal{E} (or \underline{P} or \mathcal{M}) is **finitely generated** when it is the natural extension of some finite class \mathcal{D} of simple gambles.¹ (A gamble is **simple** when it has only finitely many possible values.) In that case, the subfield of events generated by all events of the form $\{\omega: X(\omega) = \mu\}$, for real μ and X in \mathcal{D} , is finite. Since all gambles in \mathcal{D} are measurable with respect to this subfield, for purposes of computing \mathcal{E} , \underline{P} and \mathcal{M} we can replace Ω by the finite set of atoms of the subfield. (The model can then be refined to the original space Ω as described in section 4.3.3.) That means that Ω is effectively finite for finitely generated models, and, together with finiteness of \mathcal{D} , this simplifies the computation of \mathcal{E} , \underline{P} and \mathcal{M} . When You assess the upper and lower probabilities of finitely many events, for example, the resulting model is finitely generated. Other finitely generated models are described in sections 4.4 to 4.6.

4.2 FINITELY GENERATED MODELS

Suppose then that both \mathcal{D} and Ω are finite. Write $\mathcal{D} = \{X_1, \dots, X_m\}$, $\Omega = \{\omega_1, \dots, \omega_k\}$, and let $X_{m+j} = \{\omega_j\}$ for $1 \leq j \leq k$. Then the gambles X_1, \dots, X_n , where $n = m + k$, must all be almost desirable. The natural extension \mathcal{E} can be written as $\mathcal{E} = \{\sum_{j=1}^n \lambda_j X_j : \lambda_j \geq 0\}$, which is the closed polyhedral cone generated by X_1, \dots, X_n .

By a basic property of polyhedral cones,² $\mathcal{M} = \mathcal{M}(\mathcal{D}) = \mathcal{M}(\mathcal{E})$ is the convex hull of a finite set \mathcal{M}_0 of linear previsions, and $\mathcal{E} = \{X \in \mathcal{L}: P(X) \geq 0 \text{ for all } P \in \mathcal{M}_0\}$. Moreover, \mathcal{M}_0 can be taken to be the set $\text{ext } \mathcal{M}$ of extreme points of \mathcal{M} , each of which is the unique element of \mathcal{M} satisfying some maximal consistent subset of the equations $P(X_j) = 0$. Thus each extreme point P of \mathcal{M} is the unique solution of a finite system of linear inequalities $P(X_j) \geq 0$ for $1 \leq j \leq n$ and $P(1) = 1$, in which there is equality for some maximal subset of the n inequalities. This system has solutions just when \mathcal{D} avoids sure loss.³ The solutions can be found by standard linear programming algorithms.⁴

The lower prevision \underline{P} is defined for all gambles X by $\underline{P}(X) = \max \{\mu: X - \mu \in \mathcal{E}\} = \max \{\mu: X - \mu \geq \sum_{j=1}^n \lambda_j X_j, \lambda_j \geq 0, X_j \in \mathcal{D}\}$. Thus $\underline{P}(X)$ may be obtained by maximizing μ subject to the linear constraints $\mu \leq X(\omega) - \sum_{j=1}^n \lambda_j X_j(\omega)$ for all $\omega \in \Omega$ and $\lambda_j \geq 0$, another linear programming problem. The dual problem is $\underline{P}(X) = \min \{P(X): P \in \text{ext } \mathcal{M}\}$, which is a simple minimization over a finite set of real numbers if $\text{ext } \mathcal{M}$ has been computed.

It will usually be most convenient to store the model in the form of the finite set of extreme points $\text{ext } \mathcal{M}$, except perhaps when this set is much larger than \mathcal{D} . Of course, it may not be feasible to use linear programming techniques to compute $\text{ext } \mathcal{M}$ and \underline{P} when Ω or \mathcal{D} is a large set. It will then be necessary to impose some structure on \mathcal{D} to simplify the problem, as illustrated in sections 4.4 to 4.6.

4.2.2 Football example

Consider Example 4.1.4, with beliefs about $\Omega = \{W, D, L\}$ elicited through $\mathcal{D} = \{X_1, X_2, X_3, X_4\}$. This is a finitely generated model since both Ω and \mathcal{D} are finite. Write $X_5 = W$, $X_6 = D$, $X_7 = L$.

To find the extreme points of \mathcal{M} , consider the seven inequalities $P(X_j) \geq 0$. Each extreme point satisfies these inequalities and some maximal set of equations $P(X_j) = 0$. Now it is easy to see that each of $P(W) = 0$, $P(D) = 0$, $P(L) = 0$ is inconsistent with $P(X_j) \geq 0$ for $1 \leq j \leq 4$, so we need only consider $P(X_j) = 0$ for $X_j \in \mathcal{D}$. Any pair X_i, X_j of the gambles in \mathcal{D} determines a unique additive P , by solving $P(X_i) = P(X_j) = 0$ and $P(1) = 1$. The six probability mass functions $P = (P(W), P(D), P(L))$ determined in this way are $P_1 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ from X_2 and X_3 ; $P_2 = (0.5, 0.25, 0.25)$ from X_1 and X_3 ; $P_3 = (0.5, 0.3, 0.2)$ from X_1 and X_4 ; $P_4 = (0.4, 0.4, 0.2)$ from X_2 and X_4 .

$P_5 = (0.6, 0.2, 0.2)$ from X_3 and X_4 ; $P_6 = (0.5, 0.5, 0)$ from X_1 and X_2 . Of these six linear previsions, the first four satisfy all seven inequalities $P(X_j) \geq 0$, but P_5 and P_6 do not, since $P_5(X_1) = -0.2 = P_6(X_4)$. Hence the extreme points of \mathcal{M} are just P_1, P_2, P_3, P_4 .

It is then easy to compute $P(X)$ or $\bar{P}(X)$ for any gamble X by minimizing or maximizing $P_j(X)$ over $1 \leq j \leq 4$. For example, $\underline{P}(D) = 0.25$ (achieved by P_2) and $\bar{P}(D) = 0.4$ (achieved by P_4).

4.2.3 Simplex representation⁵

Linear previsions on a three-point space Ω can be conveniently represented as points in the two-dimensional **probability simplex**. This is an equilateral triangle with height one unit, in which the probabilities assigned to the three elements of Ω are identified with perpendicular distances from the three sides of the triangle. The simplex representation is especially useful for constructing \mathcal{M} and for studying the effects of new judgements of desirability.

Gambles X can be identified with their corresponding hyperplanes $\{P \in \mathcal{P}: P(X) = 0\}$, which are straight lines cutting the simplex. Since \mathcal{M} is just the intersection over $X_j \in \mathcal{D}$ of the half-spaces $\mathcal{M}_j = \{P \in \mathcal{P}: P(X_j) \geq 0\}$ bounded by such lines, \mathcal{M} can be constructed graphically by plotting the line corresponding to each gamble in \mathcal{D} . This has been done in Figure 4.2.3 for the set $\mathcal{D} = \{X_1, X_2, X_3, X_4\}$ from Example 4.2.2. The closed convex polyhedron \mathcal{M} is the shaded region bounded by the four lines, and its four extreme points are intersection points of pairs of lines. Judgements of upper and lower probabilities of singletons (X_1 and X_4) correspond to lines parallel to sides of the triangle. The comparative probability judgements X_2 and X_3 are represented by lines bisecting the triangle.

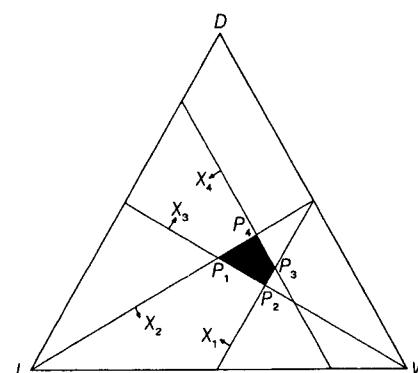


Figure 4.2.3 Simplex representation for Example 4.2.2

4.3 STEPS IN THE ASSESSMENT PROCESS

The class \mathcal{E} of almost-desirable gambles consists of all gambles whose corresponding half-space contains \mathcal{M} . The undesirable gambles are those whose half-space does not intersect \mathcal{M} . Desirability is indeterminate for the other gambles, which are represented by lines cutting \mathcal{M} . In the football example, the gamble $D - 0.2$ is desirable, $D - 0.5$ is undesirable, and $D - 0.3$ is indeterminate.

The upper and lower probabilities of singletons can be read off the simplex representation by plotting lines that are tangent to \mathcal{M} and parallel to the sides of the simplex. For example, the lines $\underline{P}(D) = 0.25$ and $\bar{P}(D) = 0.4$ are tangent to \mathcal{M} , giving $\underline{P}(D) = 0.25$ and $\bar{P}(D) = 0.4$. More generally, $\underline{P}(X)$ and $\bar{P}(X)$ are just the values of μ such that $P(X) = \mu$ is tangent to \mathcal{M} .

This example again illustrates that the lower prevision \underline{P} is not determined by the upper and lower probabilities of events. If You assess only these upper and lower probabilities, the natural extension of the assessments, \mathcal{M}' , is larger than \mathcal{M} . (In fact \mathcal{M}' is the smallest polygon whose sides are parallel to the three sides of the simplex and which contains \mathcal{M} .)

When the judgements \mathcal{D} incur sure loss, the construction produces an empty set \mathcal{M} , and we can find a minimal set of gambles in \mathcal{D} which incur sure loss by identifying a minimal set of half-spaces whose intersection is empty. Suppose that the gamble $X_5 = 0.2 - D$ is added to \mathcal{D} in the football example. The half-space \mathcal{M}_5 determined by X_5 does not intersect \mathcal{M} , indicating that the extended set of desirable gambles incurs sure loss. The intersection of half-spaces $\mathcal{M}_1, \mathcal{M}_3$ and \mathcal{M}_5 is empty, indicating that the judgements corresponding to X_1, X_3, X_5 are inconsistent. (In fact, $X_1 + X_3 + 2X_5 = -0.1$.)

More generally, the effect of a further judgement that X is almost desirable can be seen by intersecting \mathcal{M} with the new half-space \mathcal{M}_x . The simplex representation can be used also to display the effects of other kinds of judgements (described in section 4.3) and to compare the probabilities of various assessors (Appendix I3).

4.3 Steps in the assessment process

Probability assessments are not made instantaneously. They may need long deliberation, in which You analyse the evidence and examine the implications of possible assessments. It is therefore natural to view assessment and elicitation as processes which develop in time. This also allows the assessment process to be seen in the wider context of the evolution of beliefs, which includes related processes such as formulation and modification of a possibility space, re-analysis of evidence using new assessment strategies, and modification of beliefs in the light of new evidence.

One advantage of viewing assessment as a sequential process is that it allows the currently elicited model to influence later assessment. Analysis of the current model can show which uncertainties need to be assessed most carefully. In inference problems, You might concentrate on assessing probabilities most precisely for those uncertainties which contribute most to the indeterminacy in conclusions. In a decision problem, elicitation may be guided by the aim of progressively reducing the set of maximal actions.

During this process it is necessary to weigh the likely benefits from further analysis against the effort this requires, to decide how much effort should be devoted to assessment and to what uncertainties it should be directed. In particular, it will sometimes be unprofitable to narrow down intervals $[\underline{P}(A), \bar{P}(A)]$ by careful assessment, even when that is possible, because of the amount of hard thinking required or because a precise evaluation is not needed in the problem at hand.

Further analysis will usually tend to increase the precision of the probability model, by enlarging \mathcal{E} , increasing \underline{P} , and reducing \mathcal{M} and the set of maximal actions (section 4.3.1). Sometimes further analysis will have the opposite effect. That may happen when a new judgement is found to be inconsistent with earlier judgements, so that the enlarged set \mathcal{D} incurs sure loss, and this is resolved by revoking some of the earlier judgements (section 4.3.2). Even when \mathcal{D} avoids sure loss, it may be that the implications of Your judgements, revealed to You through \mathcal{E} , \underline{P} and \mathcal{M} , are unacceptable to You and lead You to modify some of the judgements. Indeed, all judgements made in the assessment process should be subjected to continual reappraisal as more is learned about their interaction with other judgements.¹

Suppose that, at the current stage of assessment, Your direct judgements of desirability are represented by \mathcal{D}_0 , and its natural extensions are \mathcal{E}_0 , P_0 , \mathcal{M}_0 . (We assume that these models are finitely generated.) The next step in assessment will modify this model in some way to give a new model \mathcal{D}_1 , \mathcal{E}_1 , \underline{P}_1 , \mathcal{M}_1 . In the following subsections we consider some simple kinds of modification and describe the changes induced in the models \mathcal{E} , \underline{P} , \mathcal{M} .² We will illustrate these changes using the football example, for which $\mathcal{D}_0 = \{X_1, X_2, X_3, X_4\}$ and $\text{ext } \mathcal{M}_0 = \{P_1, P_2, P_3, P_4\}$ are defined in Example 4.1.4.

4.3.1 Enlarging \mathcal{D}

Suppose You make a further direct judgement, with the effect of adding gambles to \mathcal{D} . For simplicity, suppose that a single gamble Z is added to \mathcal{D}_0 , giving $\mathcal{D}_1 = \mathcal{D}_0 \cup \{Z\}$. Obviously \mathcal{E}_1 will contain \mathcal{E}_0 , and \mathcal{E}_1 strictly

contains \mathcal{E}_0 unless Z is already in \mathcal{E}_0 , so the new judgement increases the precision of the model.

In computing the new model the first step is to check whether \mathcal{D}_1 avoids sure loss. Assuming that \mathcal{D}_0 avoids sure loss, the following conditions are equivalent.

- (a) \mathcal{D}_1 avoids sure loss
- (b) $-Z$ is not in the interior of \mathcal{E}_0
- (c) $\bar{P}_0(Z) \geq 0$
- (d) $P(Z) \geq 0$ for some P in $\text{ext } \mathcal{M}_0$
- (e) $\mathcal{M}_0 \cap \mathcal{M}_z$ is non-empty, where $\mathcal{M}_z = \{P \in \mathcal{P} : P(Z) \geq 0\}$ is the half-space corresponding to Z .

It is usually simplest to check condition (d). Suppose, in the football example 4.1.4, You judge that win is at least 1.5 times as probable as loss, so that $Z = W - 1.5L$ is added to \mathcal{D}_0 . Because $P_3(Z) = 0.2$, condition (d) holds, and the extended set of judgements \mathcal{D}_1 avoids sure loss.

Alternatively, condition (e) can be checked using the simplex representation 4.2.3, by plotting the line $P(W) = 1.5P(L)$ which represents Z . This is shown in Figure 4.3.1. The half-space \mathcal{M}_z , which lies to the right of this line, does intersect \mathcal{M}_0 . The intersection $\mathcal{M}_0 \cap \mathcal{M}_z$ is actually the natural extension \mathcal{M}_1 of the extended set of judgements. This is the shaded region in the figure.

The extreme points of the new model \mathcal{M}_1 consist of the extreme points of \mathcal{M}_0 which lie in the half-space \mathcal{M}_z , together with the points at which the hyperplane $P(Z) = 0$ intersects the faces of \mathcal{M}_0 . In the football example, \mathcal{M}_1 has five extreme points: the old extreme points P_2, P_3, P_4 , which lie in \mathcal{M}_z , and the new extreme points $Q_1 = (\frac{3}{8}, \frac{3}{8}, \frac{1}{4})$ and $Q_2 = (\frac{3}{7}, \frac{2}{7}, \frac{2}{7})$, at which the line Z intersects the lines X_2 and X_3 (sides of \mathcal{M}_0).

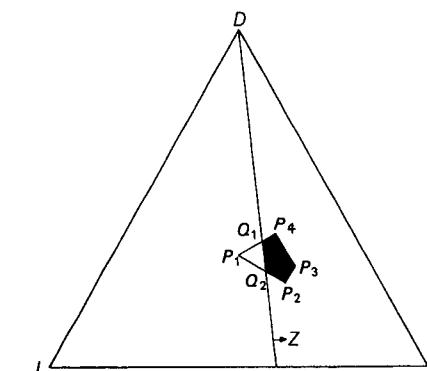


Figure 4.3.1 New model produced by an extra judgement

For finitely generated models, the new extreme points can be computed by solving $P(Z) = 0$ plus a maximal set of equations $P(X_j) = 0$ from the old model. The new lower prevision \underline{P}_1 can then be obtained by minimization over the finite set of extreme points of \mathcal{M}_1 .³ In the football example, the lower and upper probabilities for win (0.375, 0.5) and loss (0.2, 0.286) are made more precise by the extra judgement Z , while the probabilities for draw (0.25, 0.4) are unchanged.

4.3.2 Reducing \mathcal{D}

The natural extension of \mathcal{D}_0 summarizes the implications of Your direct judgements. After examining these implications You may decide that some of them are unacceptable, and You may then wish to revoke some of the direct judgements. (That is necessary, of course, when \mathcal{D}_0 incurs sure loss.)

Suppose that \mathcal{D}_1 is constructed by removing a single gamble Z from \mathcal{D}_0 . Then \mathcal{M}_1 consists of \mathcal{M}_0 plus some subset of the half-space $\mathcal{M}_Z^c = \{P \in \mathcal{P}: P(Z) < 0\}$. All the extreme points of \mathcal{M}_0 which satisfy $P(Z) > 0$ are extreme points of \mathcal{M}_1 . If the extreme points of \mathcal{M}_0 have been found by solving maximal consistent sets of equations, as in Example 4.2.2, it is quite easy to obtain the new extreme points of \mathcal{M}_1 , which are determined by sets of equations not involving Z .

Suppose that the first judgement is revoked in the football example 4.1.4, so that $Z = X_1 = 0.5 - W$. Then the three extreme points of \mathcal{M}_1 are P_1 and P_4 , which are the extreme points of \mathcal{M}_0 for which $P(Z) = 0.5 - P(W) > 0$, and $P_5 = (0.6, 0.2, 0.2)$, which is the intersection point of the lines X_3 and X_4 but satisfies $P_5(Z) < 0$.⁴ The triangle \mathcal{M}_1 can be easily identified in the probability simplex displayed in Figure 4.2.3.

4.3.3 Refining Ω

The possibility space Ω may itself evolve during elicitation. Although the choice of Ω is not crucial in our approach, since we do not require that precise probabilities be assessed for all subsets of Ω , we do require that Ω be fine enough to represent all the probability judgements that You wish to make, so that all Your direct judgements can be identified with gambles defined on Ω . During elicitation Ω may need to be extended, either by refining some of its possible states or adding new states not previously regarded as possible. Refinement of Ω will often be necessary when You decide to analyse a problem in greater detail.⁵

Formally, refinement of Ω_0 to give a new space Ω_1 can be represented by a mapping from Ω_1 onto Ω_0 . Then Ω_0 corresponds to a partition of Ω_1 ,

4.3 STEPS IN THE ASSESSMENT PROCESS

and we can identify each state ω in Ω_0 with the set $A(\omega)$ of ‘refined possibilities’ in Ω_1 which map to ω . Often Ω_1 will be the product space of Ω_0 with another possibility space Ψ , with $A(\omega) = \{\omega\} \times \Psi$. In the football example the space $\Omega_0 = \{W, D, L\}$ might be refined to the space Ω_1 of ordered pairs (i, j) of non-negative integers, representing the possible scores for each team, with $A(W) = \{(i, j): i > j\}$, $A(D) = \{(i, j): i = j\}$, $A(L) = \{(i, j): i < j\}$.

How does refinement of Ω_0 change the model \mathcal{E}_0 , \underline{P}_0 , \mathcal{M}_0 defined on $\mathcal{L}(\Omega_0)$? We can identify $\mathcal{L}(\Omega_0)$ with the linear subspace \mathcal{K} of $\mathcal{L}(\Omega_1)$ consisting of those gambles on Ω_1 that are constant on $A(\omega)$ for each $\omega \in \Omega_0$. The model \mathcal{E}_0 provides evaluations of just the gambles in \mathcal{K} , and this can be extended to $\mathcal{L}(\Omega_1)$ by natural extension.

Formally, we can identify each gamble Y in $\mathcal{L}(\Omega_0)$ with \tilde{Y} in $\mathcal{L}(\Omega_1)$, defined by $\tilde{Y}(v) = Y(\omega)$ whenever $v \in A(\omega)$. For X in $\mathcal{L}(\Omega_1)$, define the lower bound X_* in $\mathcal{L}(\Omega_0)$ by $X_*(\omega) = \inf\{X(v): v \in A(\omega)\}$. Using Corollary 3.1.8, the refined models are $\underline{P}_1(X) = \underline{P}_0(X_*)$, $\mathcal{E}_1 = \{X \in \mathcal{L}(\Omega_1): X_* \in \mathcal{E}_0\}$, and $\mathcal{M}_1 = \{P \in \mathcal{P}(\Omega_1): P^0 \in \mathcal{M}_0\}$, where P^0 is the **marginal** of P , defined by $P^0(Y) = P(\tilde{Y})$. The extreme points of \mathcal{M}_1 can be obtained from the extreme points P of \mathcal{M}_0 , by assigning all the probability $P(\{\omega\})$ to a single point of $A(\omega)$, for each ω in Ω_0 .

As a simple example, suppose You refine the space $\Omega_0 = \{W, D, L\}$ in the football example by splitting the outcome ‘draw’ into the two possibilities ‘scoring draw’ (S) and ‘goalless draw’ (G).⁶ Then $\Omega_1 = \{W, S, G, L\}$, and the partition A is $A(W) = \{W\}$, $A(D) = \{S, G\}$, $A(L) = \{L\}$. The eight extreme points of \mathcal{M}_1 are obtained from the four extreme points in Example 4.1.4 by assigning all of the probability $P_j(D)$ either to S or to G .

The formal step of refining Ω does not in itself provide new information about beliefs, but merely represents the old information in a different form. In the example it is clear that the new model can contain no information to distinguish S from G , because such information could not have been represented in terms of Ω_0 . The purpose of refining Ω_0 is to enable this information, when elicited, to be represented.

4.3.4 Coarsening and marginalization

The opposite transformation to refinement is called **coarsening**. Formally, a coarsening is a mapping A from Ω_1 onto Ω_0 . Since the elements v of Ω_1 correspond to disjoint subsets $A^{-1}(v)$ of Ω_1 , the coarser space Ω_1 can be identified with a partition of Ω_0 . For X in $\mathcal{L}(\Omega_1)$, define X_* in $\mathcal{L}(\Omega_0)$ by $X_*(\omega) = X(A(\omega))$. For P in $\mathcal{P}(\Omega_1)$, define the **marginal** P^1 in $\mathcal{P}(\Omega_1)$ by $P^1(X) = P(X_*)$. Then the new model can be obtained from the old by $\underline{P}_1(X) = \underline{P}_0(X_*)$, $\mathcal{E}_1 = \{X \in \mathcal{L}(\Omega_1): X_* \in \mathcal{E}_0\}$ and $\mathcal{M}_1 = \{P^1: P \in \mathcal{M}_0\}$. The

extreme points of \mathcal{M}_1 are a subset of $\{P^1: P \in \text{ext } \mathcal{M}_0\}$, the marginals of extreme points of \mathcal{M}_0 .

In the football example, if You are interested only in whether the game is drawn then You may coarsen Ω_0 to $\Omega_1 = \{D, R\}$ by identifying R (for ‘result’) with $\{W, L\}$. The assessments in Example 4.1.4 yield $\underline{P}_1(R) = \underline{P}_0(W + L) = 0.6$, $\bar{P}_1(R) = 0.75$, $\mathcal{M}_1 = \{P \in \mathcal{P}(\Omega_1): 0.6 \leq P(R) \leq 0.75\}$. The two extreme points of \mathcal{M}_1 are $(0.25, 0.75)$ and $(0.4, 0.6)$, the marginals of P_2 and P_4 . In terms of the simplex representation 4.2.3, \mathcal{M}_1 is the orthogonal projection of \mathcal{M}_0 onto the line through D which bisects the probability triangle. (This line is the one-dimensional simplex representing $\mathcal{P}(\Omega_1)$.)

The special type of coarsening where $\Omega_0 = \Omega_1 \times \Psi$ is a product space and $A^{-1}(v) = \{v\} \times \Psi$ is known as **marginalization**. When B is a subset of Ω_1 , $B_* = B \times \Psi$, hence $\underline{P}_1(B) = \underline{P}_0(B \times \Psi)$ and $P^1(B) = P(B \times \Psi)$. The marginal probabilities of sets B are just the probabilities of the corresponding sets $B \times \Psi$ under the original model.

4.3.5 Multivalued mappings

To generalize the refinement and coarsening transformations, suppose that each state ω in Ω_0 is consistent with each state in some non-empty subset $A(\omega)$ of the new space Ω_1 . Knowing ω tells You that the true state of Ω_1 belongs to $A(\omega)$, but tells You nothing more about this state. Dempster (1967a) calls A a **multivalued mapping** from Ω_0 to Ω_1 . Suppose that every state of Ω_1 is contained in some $A(\omega)$. When all the sets $A(\omega)$ are disjoint, Ω_1 is a **refinement** of Ω_0 . When each $A(\omega)$ is a singleton set, Ω_1 is a **coarsening** of Ω_0 . More generally, the sets $A(\omega)$ may contain many states and may intersect.

Define X_* as in section 4.3.3, to be the smallest gamble on Ω_0 that is consistent with the gamble X on Ω_1 . When ω occurs You will get at least $X_*(\omega)$ from the gamble X . From the model \mathcal{E}_0 , \underline{P}_0 , \mathcal{M}_0 defined on Ω_0 , we can conclude that X is desirable (in \mathcal{E}_1) only when X_* is desirable (in \mathcal{E}_0). Hence the multivalued mapping induces the new models $\mathcal{E}_1 = \{X \in \mathcal{L}(\Omega_1): X_* \in \mathcal{E}_0\}$ and $\underline{P}_1(X) = \underline{P}_0(X_*)$, as in the special cases of refinement and coarsening. Provided Ω_0 and Ω_1 are both finite, the linear previsions in \mathcal{M}_1 can be obtained from those P in \mathcal{M}_0 by distributing each mass $P(\{\omega\})$ arbitrarily amongst the elements of $A(\omega)$. The extreme points of \mathcal{M}_1 can be obtained from those of \mathcal{M}_0 by assigning each $P(\{\omega\})$ to a single element of $A(\omega)$.⁷

In the football example with $\Omega_0 = \{W, D, L\}$, interest might be focused on the total number of goals scored in the game. Let $\Omega_1 = \{N, O, M\}$, where N represents ‘no goals’, O is ‘one goal’ and M is ‘more than one goal’. The associated multivalued mapping is $A(W) = A(L) = \{O, M\}$ and $A(D) = \{N, M\}$. Here Ω_1 is neither a refinement nor a coarsening of Ω_0 .

4.3 STEPS IN THE ASSESSMENT PROCESS

The extreme points of \mathcal{M}_1 can be found from the four extreme points P_j of \mathcal{M}_0 , by assigning the probability $P_j(D)$ to either N or M and assigning $1 - P_j(D)$ to either O or M . Only P_2 and P_4 , which achieve $\underline{P}_0(D)$ and $\bar{P}_0(D)$, are relevant, and they give the five extreme points $(P(N), P(O), P(M)) = (0, 0, 1), (0.4, 0.6, 0), (0.4, 0, 0.6), (0.25, 0.75, 0)$ and $(0, 0.75, 0.25)$. The simplex representation in Figure 4.3.5a shows that \mathcal{M}_1 is large; the beliefs about Ω_0 are relatively uninformative about Ω_1 . Here \mathcal{M}_1 is the intersection of the two half-spaces $P(O) + P(M) \geq 0.6$ and $P(N) + P(M) \geq 0.25$, which are generated by the assessments $\underline{P}_0(W \cup L) = 0.6$ and $\bar{P}_0(D) = 0.25$ through the multivalued mapping.⁸

Now consider the special case where Ω_0 is finite and the initial model P_0 is a linear prevision. When B is a subset of Ω_1 , $\underline{P}_1(B) = P_0(B_*) = P_0(\{\omega: B \supset A(\omega)\})$. In this case the lower probability \underline{P}_1 is called a **belief function**.⁹

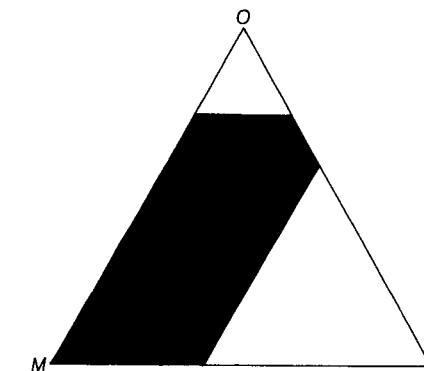


Figure 4.3.5a New model produced by a multivalued mapping

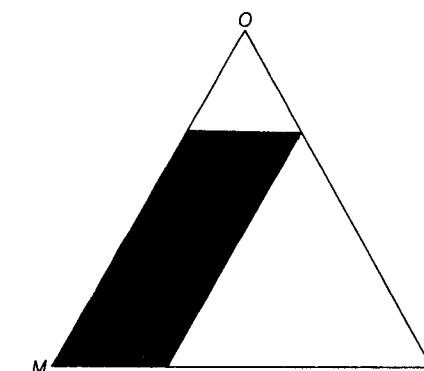


Figure 4.3.5b New model produced by a multivalued mapping and precise initial probabilities

Consider the football example with the initial mass function $P_0 = (0.5, 0.3, 0.2)$ and the multivalued mapping defined above. The resulting class \mathcal{M}_1 (shown in the Figure 4.3.5b) is the intersection of the two half-spaces $P(O) + P(M) \geq 0.7$ and $P(N) + P(M) \geq 0.3$. The corresponding belief function is defined by $\underline{P}_1(\Omega_1) = 1$, $\underline{P}_1(\{O, M\}) = 0.7$, $\underline{P}_1(\{N, M\}) = 0.3$, $\underline{P}_1(B) = 0$ for all other subsets B . The new model is highly imprecise, despite the precision of P_0 .

By the preceding argument, Your beliefs concerning Ω_1 can be represented by a belief function \underline{P}_1 when three conditions are satisfied: (a) Ω_1 is related to Ω_0 through a multivalued mapping, (b) Your probabilities concerning Ω_0 are precise, and (c) You make no other judgements of probabilities on Ω_1 . Each of these conditions is restrictive. Multivalued mappings are one amongst many ways of constructing probability models, precise probabilities are not usually justified, and there may be additional information about Ω_1 that can be used to increase the precision of \underline{P}_1 .¹⁰ In the football example, it is reasonable to sharpen the belief function \underline{P}_1 by making a further assessment that $\bar{P}(M) \leq 0.9$, but then the new model \underline{P}_2 is not a belief function.

4.3.6 Enlarging Ω

During assessment You may decide that Your current space Ω_0 is not an exhaustive set of possibilities. You may admit a set Ψ of new possibilities, so enlarging Ω_0 to $\Omega_1 = \Omega_0 \cup \Psi$. Such a change could be motivated by a more careful analysis of a problem, by empirical observation of states not previously considered possible, or by statistical data that suggest new hypotheses.

Gambles X in $\mathcal{L}(\Omega_0)$ can be identified with gambles X' in $\mathcal{L}(\Omega_1)$ which agree with X on Ω_0 , but are null (or ‘called off’) if Ω_0 fails to occur, so $X'(\omega) = X(\omega)$ if $\omega \in \Omega_0$, $X'(\omega) = 0$ if $\omega \in \Psi$. We take X' to be desirable just when X is desirable. For Y in $\mathcal{L}(\Omega_1)$, let Y_0 and Y_1 denote the restrictions to Ω_0 and Ψ respectively. The new model is then defined by $\mathcal{E}_1 = \{Y \in \mathcal{L}(\Omega_1): Y_0 \in \mathcal{E}_0, Y_1 \geq 0\}$, $\underline{P}_1(Y) = \min\{\underline{P}_0(Y_0), \inf Y_1\}$, and \mathcal{M}_1 consists of all linear previsions of the form $P(Y) = \lambda P_0(Y_0) + (1 - \lambda)P_1(Y_1)$, which are convex combinations of previsions P_0 in \mathcal{M}_0 with arbitrary previsions P_1 on Ψ . The extreme points of \mathcal{M}_1 are just the extreme points P_0 of \mathcal{M}_0 , extended by taking $P_0(\Psi) = 0$, together with the extreme points P_1 of $\mathcal{P}(\Psi)$, extended by $P_1(\Omega_0) = 0$.¹¹ The new model assigns vacuous probabilities to Ψ . Of course, You would typically judge $\bar{P}(\Psi)$ to be small, but that requires further assessment and is not simply a consequence of enlarging Ω_0 .

4.3 STEPS IN THE ASSESSMENT PROCESS

In the football example You might enlarge Ω_0 to include the possibility (A) that the game is abandoned without a result. Then the five extreme points of \mathcal{M}_1 are the degenerate prevision $P(A) = 1$, plus the four extreme points P_j of \mathcal{M}_0 extended by $P_j(A) = 0$. It would be sensible to assess a small value for $\bar{P}(A)$.

4.3.7 Reducing Ω

As suggested in section 2.1.2, You may wish to reduce Ω_0 to a proper subset Ω_1 when analysis of the available evidence reveals that some states which previously appeared possible are logically or practically inconsistent with the evidence. In some problems You may reduce Ω_0 to a set of pragmatic possibilities by removing states that You consider practically possible but highly improbable or otherwise negligible. You may be satisfied with a simpler analysis based on Ω_1 . In other cases, new evidence might establish that the true state ω is in the subset Ω_1 . This is the familiar and important case of **conditioning**. In the football example You might learn that the match in question is a cup-tie for which a result (W or L) must be reached. You would then reduce Ω_0 to $\Omega_1 = \{W, L\}$.

As in section 4.3.6 we can identify gambles X in $\mathcal{L}(\Omega_1)$ with gambles X' in $\mathcal{L}(\Omega_0)$ that agree with X on Ω_1 and are zero otherwise. We will assume that the desirability of X' is not influenced by the new information that leads You to reduce Ω_0 . This assumption will be reasonable when the reduction of Ω_0 results from an attempt to simplify the analysis, rather than from new evidence, and in some cases of conditioning on new evidence. (It is unreasonable in many problems; learning that the football game is a cup-tie may well change Your judgement about which team is more likely to win.)

Assume also that $P_0(\Omega_1) > 0$. Under these assumptions, X should be desirable just when X' is. Hence the new models are $\mathcal{E}_1 = \{X \in \mathcal{L}(\Omega_1): X' \in \mathcal{E}_0\}$ and $\underline{P}_1(X) = \max\{\mu: X - \mu \in \mathcal{E}_1\} = \max\{\mu: \Omega_1(X' - \mu) \in \mathcal{E}_0\} = \max\{\mu: P_0(\Omega_1(X' - \mu)) \geq 0\}$. The last formula defines the prevision of X' conditional on Ω_1 , written as $\underline{P}_0(X'|\Omega_1)$, according to the **generalized Bayes’ rule** discussed in Chapter 6. This method of updating the model is therefore equivalent to conditioning on Ω_1 using the generalized Bayes’ rule.

The linear previsions P' in \mathcal{M}_1 can be obtained by conditioning all those P in \mathcal{M}_0 , using **Bayes’ rule** $P'(X) = P(X')/P(\Omega_1)$. The extreme points of \mathcal{M}_1 are a subset of the conditioned extreme points of \mathcal{M}_0 .

In the football example with $\Omega_1 = \{W, L\}$, we obtain the two extreme points of \mathcal{M}_1 by conditioning P_1 and P_3 , giving $(P(W), P(L)) = (0.5, 0.5)$ and $(\frac{5}{7}, \frac{2}{7})$. The new model is determined by the upper and lower probabilities of W , which are $\frac{5}{7}$ and $\frac{1}{2}$.

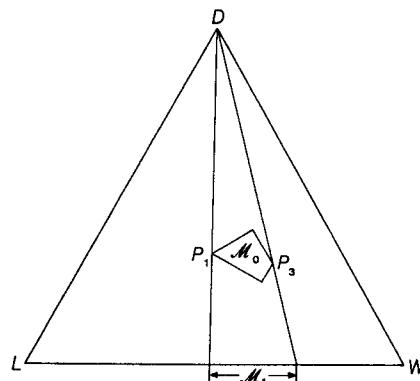


Figure 4.3.7 *New model produced by conditioning*

The old and new models, \mathcal{M}_0 and \mathcal{M}_1 , are represented in the probability simplex in Figure 4.3.7. A gamble X on Ω_1 is represented by a point on the line LW , and the corresponding gamble X' on Ω_0 is represented by the line joining this point to the vertex D . Similarly \mathcal{M}_1 is the projection of \mathcal{M}_0 from the vertex D onto the line LW .

4.3.8 Combining assessments

Suppose we have two sources of information about the possibility space Ω , which are represented by the two models \mathcal{E}_j , P_j , \mathcal{M}_j for $j = 1, 2$. The two models might be the assessments of a single person using two different elicitation methods or assessment strategies, or they might be the judgements of two different people. How should the two probability models be combined?

For example, the two models might refer to logically independent possibility spaces Ω_1 and Ω_2 , e.g. the outcomes of two football games. Both models can be represented on the product space $\Omega = \Omega_1 \times \Omega_2$ using the refinement technique of section 4.3.3. They can then be aggregated into a single model on Ω .

When the two models are **consistent**, in the sense that the combined class of desirable gambles $\mathcal{E}_1 \cup \mathcal{E}_2$ avoids sure loss, they can be regarded as sources of information about a single belief-state (of an individual or group). It is then natural to regard all the gambles in $\mathcal{E}_1 \cup \mathcal{E}_2$ as almost desirable, so that the combined model \mathcal{E} is the natural extension (convex hull) of $\mathcal{E}_1 \cup \mathcal{E}_2$. The combined lower prevision \underline{P} is the natural extension of \underline{Q} where $\underline{Q}(X) = \max\{\underline{P}_1(X), \underline{P}_2(X)\}$, and the combined class of linear previsions \mathcal{M} is the intersection $\mathcal{M}_1 \cap \mathcal{M}_2$, which is non-empty just when the two models are consistent. This rule of combination is called the **conjunction rule**.¹²

4.3 STEPS IN THE ASSESSMENT PROCESS

In the case where the two models refer to different spaces Ω_1 and Ω_2 , refinement to the product space $\Omega = \Omega_1 \times \Omega_2$ gives the models \mathcal{M}_1 and \mathcal{M}'_2 , where \mathcal{M}'_j consists of all linear previsions on Ω whose Ω_j -marginal is in \mathcal{M}_j . The conjunction rule gives combined model $\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}'_2$ which consists of all previsions with Ω_1 -marginal in \mathcal{M}_1 and Ω_2 -marginal in \mathcal{M}'_2 . In this case the two models are always consistent and \mathcal{M} is always non-empty. A precise model could be obtained by judging that the two marginals are epistemically independent (see section 9.3).

It is clearly beneficial to use several assessment strategies or elicitation methods, provided they give consistent results, because the combined model is more precise than its component models.¹³ (The exception is when \mathcal{M}_1 is contained in \mathcal{M}_2 , so the second model adds no information to the first.) The benefits will be greatest when the two models focus on different aspects of Ω , as in the example of the previous paragraph.

4.3.9 Inconsistent assessments

We would expect consistency of the two probability models when they result from different assessments of the same body of evidence, or when they are based on unrelated bodies of evidence referring to different possibility spaces, e.g. when they are obtained from two experts whose areas of expertise do not overlap. There is less reason to expect consistency when the two bodies of evidence partially overlap or conflict.¹⁴

The two models are inconsistent when \mathcal{M}_1 and \mathcal{M}_2 are disjoint. In that case there is a hyperplane which strongly separates them (Appendix E3). This corresponds to a gamble X for which $\underline{P}_1(X)$ is positive and $\underline{P}_2(X)$ is negative, so X is strictly desirable under the first model, and $-X$ is strictly desirable under the second. If the two models are elicited from different people, then the two people are willing to bet with each other by exchanging the gamble X . (When their models are consistent, they are not willing to exchange any gamble.)

Consider the football example. Suppose that one assessor makes the judgements in Example 4.1.4, giving model \mathcal{M}_1 . A second assessor judges that $\underline{P}(W) \leq 0.4$, $\underline{P}(L) \leq 0.4$ and W is at least 1.5 times as probable as D , yielding the model \mathcal{M}_2 determined by the desirable gambles $0.4 - W$, $0.4 - L$ and $W - 1.5D$. The two polygons \mathcal{M}_1 and \mathcal{M}_2 are shown in the probability simplex in Figure 4.3.9. They are disjoint, indicating that the two models are inconsistent, and they can be separated by the hyperplane $X = D - 0.9L$.

When the two models are inconsistent, it is advisable to investigate the source of the inconsistency before combining them. If there is conflict between the bodies of evidence on which the two models are based then it is reasonable for the combined model to be less precise than the component models.

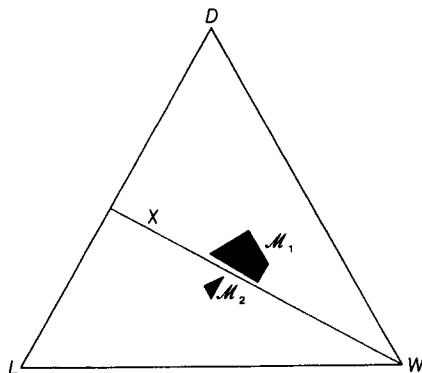


Figure 4.3.9 Inconsistent probability models

The conjunction rule is inapplicable in cases of inconsistency. A simple alternative is the **unanimity rule**, which takes a gamble to be desirable if and only if it is desirable under each of the component models. Thus the combined model is $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$, \underline{P} is the lower envelope $\underline{P}(X) = \min\{\underline{P}_1(X), \underline{P}_2(X)\}$, and \mathcal{M} is the convex hull of $\mathcal{M}_1 \cup \mathcal{M}_2$.

In the preceding football example, the combination of \mathcal{M}_1 and \mathcal{M}_2 by the unanimity rule can be easily found from the simplex representation. It is just the smallest convex set containing \mathcal{M}_1 and \mathcal{M}_2 . It has six extreme points: the four extreme points of \mathcal{M}_1 , and two extreme points (0.36, 0.24, 0.4) and (0.4, 0.2, 0.4) of \mathcal{M}_2 .

The combined model produced by the unanimity rule is always coherent, but it will often be highly imprecise. Its imprecision is at least as large as the imprecision of the component models, and also reflects the extent of disagreement between them.¹⁵

4.4 Classificatory probability

In the rest of this chapter we examine models that are constructed from special types of probability judgements. The simplest type of judgement, and the most common in ordinary life, is that some event is probable. The general elicitation procedure (section 4.1) can use such qualitative judgements, and it is interesting to see how much information they can provide about beliefs. We therefore consider the simplest concept of probability, which is called **classificatory probability**.¹

4.4.1 Behavioural interpretation

As in section 4.1.1, a judgement that event B is probable is given the behavioural interpretation that You are willing to bet on B at any odds

4.4 CLASSIFICATORY PROBABILITY

better than even money.² In fact, we can bypass the difficulties in constructing a utility scale for gambles by asking You to compare a contract which pays You £100 if B occurs (and nothing otherwise) with one which pays You £100 if B fails to occur. If You prefer the first contract to the second then B is probable for You.³ Most people are capable of understanding and making such judgements.

Suppose that, during elicitation, You classify some events as probable. Let \mathcal{I} be the set of all such events. Each probable event B in \mathcal{I} corresponds to an almost-desirable gamble $B - \frac{1}{2}$, an even-money bet on B , so \mathcal{I} corresponds to the class of almost-desirable gambles $\mathcal{D} = \{B - \frac{1}{2}: B \in \mathcal{I}\}$.

4.4.2 Coherence

The basic concepts of avoiding sure loss, coherence and natural extension can be applied to the classificatory judgements \mathcal{I} by applying the definitions in section 3.7 to the corresponding class \mathcal{D} . The classification \mathcal{I} incurs sure loss just when there are finitely many probable events in \mathcal{I} such that even-money bets on these events must result in a net loss. Equivalently, \mathcal{I} avoids sure loss when, for any events B_1, B_2, \dots, B_n chosen from \mathcal{I} , it is possible that at least half of the events will occur; that is, there is some state ω in Ω such that $\sum_{j=1}^n B_j(\omega) \geq n/2$. By Lemma 3.3.2, this is equivalent to the existence of a linear prevision P such that $P(B) \geq \frac{1}{2}$ for all events B in \mathcal{I} .⁴

The **natural extensions** of any classification \mathcal{I} are defined by $\mathcal{M}(\mathcal{I}) = \{P \in \mathcal{P}: P(B) \geq \frac{1}{2} \text{ for all } B \in \mathcal{I}\}$, which is the class of all linear previsions which are consistent with the classification \mathcal{I} , and

$$\underline{P}(X) = \sup \left\{ \mu: X - \mu \geq \sum_{j=1}^n \lambda_j (B_j - \frac{1}{2}) \text{ for some } B_j \in \mathcal{I}, \lambda_j \geq 0 \right\}.$$

These summarize the implications of the classification \mathcal{I} for other gambles X . In particular, the classification may imply that other events C are probable, because $\underline{P}(C) \geq \frac{1}{2}$. Thus \mathcal{I} can be extended to a class \mathcal{I}^* of probable events, defined by $\mathcal{I}^* = \{C: \underline{P}(C) \geq \frac{1}{2}\}$.⁵ Clearly \mathcal{I}^* contains \mathcal{I} . When \mathcal{I} incurs sure loss, \mathcal{I}^* contains all subsets of Ω .

Call the classification \mathcal{I} **coherent** when its corresponding \mathcal{D} is coherent relative to the class of all even-money gambles on events (Definition 3.7.2). This means that \mathcal{I} fully classifies the subsets of Ω . The following theorem summarizes the relationships between these definitions.

4.4.3 Classificatory probability theorem

Let \mathcal{I} be a class of probable events and \mathcal{I}^* its natural extension. Each of the following conditions is equivalent to \mathcal{I} avoiding sure loss.

- (a) \emptyset is not in \mathcal{I}^*
- (b) there is an additive probability P such that $P(B) \geq \frac{1}{2}$ for all $B \in \mathcal{I}$.
Each of the following conditions is equivalent to coherence of \mathcal{I} .
- (c) \emptyset is not in \mathcal{I}^* and $\mathcal{I} = \mathcal{I}^*$
- (d) there is a coherent lower probability \underline{P} , defined on all events, such that $\mathcal{I} = \{B : \underline{P}(B) \geq \frac{1}{2}\}$
- (e) there is a non-empty class \mathcal{M} of additive probabilities such that $\mathcal{I} = \{B : P(B) \geq \frac{1}{2} \text{ for all } P \in \mathcal{M}\}$.

Thus \mathcal{I} is coherent just when it agrees with its natural extension \mathcal{I}^* and does not contain all events. Equivalently, \mathcal{I} contains just those events with lower probability at least $\frac{1}{2}$ under some coherent \underline{P} .⁶ For example, if \mathcal{I} is any non-empty filter of sets (2.9.8), then the 0–1 valued lower probability defined by $\underline{P}(A) = 1$ if $A \in \mathcal{I}$, $\underline{P}(A) = 0$ otherwise, is coherent, and it follows from (d) that \mathcal{I} is coherent as a class of probable events.

4.4.4 Basic properties

The following properties of coherent classifications \mathcal{I} can be verified using the theorem.⁷

- (i) $\Omega \in \mathcal{I}$, $\emptyset \notin \mathcal{I}$
- (ii) if $C \supset B$ and $B \in \mathcal{I}$ then $C \in \mathcal{I}$
- (iii) there are no events A, B, C in \mathcal{I} such that $A \cap B = B \cap C = C \cap A = \emptyset$
- (iv) there are no events A, B, C, D in \mathcal{I} such that $A \cap B \cap C = \emptyset$ and $D \cap (A \cup B \cup C) = \emptyset$
- (v) if A, B, C are in \mathcal{I} and $A \cap B \cap C = \emptyset$ then $(A \cap B) \cup (B \cap C) \cup (C \cap A) \in \mathcal{I}$
- (vi) if A, B, C are in \mathcal{I} and $A \cap B = \emptyset$ then $(A \cup B) \cap C \in \mathcal{I}$.

If the lower probability in (d) of the theorem can be taken to be additive, then \mathcal{I} also has the **completeness** property: for each subset B of Ω , either B or B^c is in \mathcal{I} . If Your beliefs concerning Ω can be represented by an additive probability measure then Your classification \mathcal{I} must be complete.⁸ The converse does not hold. When \mathcal{I} is complete, its natural extension \underline{P} will rarely be additive, and Your beliefs may be highly indeterminate. That is illustrated by the next example.

4.4.5 Football example

Here $\Omega = \{W, D, L\}$. A judgement that event B is probable corresponds, in the simplex representation in Figure 4.4.5, to a line $P(B) = \frac{1}{2}$ joining the mid-points of two sides of the simplex. Since only the three plotted lines can be used to construct the convex set $\mathcal{M}(\mathcal{I})$, this must consist of a subset

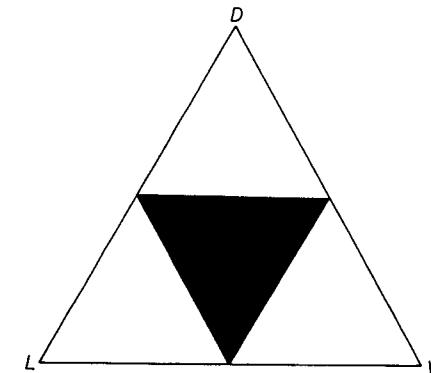


Figure 4.4.5 Model produced by classificatory judgements

of the four sub-triangles, the three plotted lines, and their three points of intersection.

For example, the classification $\mathcal{I}_1 = \{\Omega, W \cup D, D \cup L, W \cup L\}$ yields the shaded sub-triangle as its natural extension $\mathcal{M}(\mathcal{I}_1)$. Similarly, $\mathcal{I}_2 = \{\Omega, W \cup L\}$ (a filter) gives the larger region $P(D) \leq \frac{1}{2}$ as its natural extension, $\mathcal{I}_3 = \mathcal{I}_1 \cup \{D\}$ gives the line $P(D) = \frac{1}{2}$, and $\mathcal{I}_4 = \mathcal{I}_1 \cup \{W, D\}$ gives the precise point $P(W) = P(D) = \frac{1}{2}$. Each of these four classifications is coherent, all except \mathcal{I}_2 are complete, but only \mathcal{I}_4 determines a precise probability model. It is clear that classificatory judgements can provide only crude information about beliefs in this example, except in the unusual case where both an event and its complement are judged probable so that each has precise probability $\frac{1}{2}$.

Classificatory judgements can be more informative for larger spaces Ω , especially when Ω can be partitioned into many events of ‘low probability’ from which many probable events can be constructed. That can be achieved in practice by refining the possibility space of interest, for example by forming a product space with an extraneous measurement space representing the outcomes of many tosses of a fair coin (section 4.5.7). But when Ω is large there is less reason to expect completeness of \mathcal{I} .

4.5 Comparative probability

Next consider probability comparisons between events, which provide a somewhat richer structure than classificatory probability.¹ Write $A \geqslant B$ to denote a judgement that event A is at least as probable as event B . Such judgements are common in ordinary life. We suppose that You make an arbitrary set of such comparisons, yielding a partial ordering \geqslant on events which is called a **comparative probability ordering**.²

Comparative probability orderings are a special case of the almost-preference orderings studied in section 3.7. The class of almost-desirable gambles corresponding to \geqslant is $\mathcal{D} = \{A - B : A \geqslant B\}$. Thus Your judgement $A \geqslant B$ is taken to mean that You prefer any gamble that pays more than one unit of utility if A occurs (and nothing otherwise) to the gamble that pays one unit if B occurs (and nothing otherwise). Again, it is unnecessary to construct a linear utility scale. If You prefer a contract which pays You a valuable prize if A occurs (and nothing otherwise) to one which pays You the same prize if B occurs, then You judge that A is at least as probable as B . Such judgements are relatively easy to understand and elicit.

4.5.1 Coherence

The concepts of avoiding sure loss, coherence and natural extension are again applied to comparative probability orderings through the corresponding class \mathcal{D} of almost-desirable gambles. Hence the partial ordering \geqslant incurs sure loss just when You make finitely many comparisons $A_j \geqslant B_j$ such that it is certain that more B 's than A 's will occur. If You trade each gamble B_j for the gamble A_j (which You judge to be at least as good as B_j) then You are sure to make an overall loss.³ By Lemma 3.3.2, \geqslant avoids sure loss if and only if there is a linear prevision P that almost agrees with \geqslant , in the sense that $P(A) \geq P(B)$ whenever $A \geqslant B$.

The natural extension of the partial ordering, $\mathcal{M}(\geqslant)$, is the class of all almost-agreeing linear previsions, with corresponding lower prevision $\underline{P}(X) = \sup\{\mu : X - \mu \geq \sum_{j=1}^n \lambda_j(A_j - B_j) \text{ for some } A_j \geqslant B_j \text{ and } \lambda_j \geq 0\}$. The natural extension of \geqslant to a partial ordering of events, denoted by \geqslant^* , is defined by $A \geqslant^* B$ if and only if $\underline{P}(A - B) \geq 0$. Thus $A \geqslant^* B$ just when $P(A) \geq P(B)$ for all linear previsions P that almost agree with \geqslant . The natural extension \geqslant^* summarizes the implications of the judgements \geqslant for comparisons between events. Clearly $A \geqslant^* B$ whenever $A \geqslant B$, so that \geqslant^* is an extension of \geqslant .⁴

Call \geqslant **coherent** when it is coherent as a preference ordering relative to the set of all ordered pairs of events (section 3.7.5). Coherence means that no further comparisons between events can be constructed from the ordering \geqslant . Provided \geqslant avoids sure loss, its natural extension \geqslant^* is coherent, and \geqslant is coherent if and only if it agrees with \geqslant^* . The next theorem, which is analogous to Theorem 4.4.3, summarizes the conditions for coherence.

4.5.2 Comparative probability theorem

Let \geqslant be a comparative probability ordering on subsets of Ω , and \geqslant^* its natural extension. Define the relations $>$ ('is more probable than') and \approx ('is equally probable with') by $A > B$ if $A \geqslant B$ and not $B \geqslant A$, $A \approx B$ if $A \geqslant B$

4.5 COMPARATIVE PROBABILITY

and $B \geqslant A$. Define $>^*$ and \approx^* similarly. Each of the following conditions is equivalent to \geqslant avoiding sure loss.

- (a) $\Omega >^* \emptyset$
 - (b) there is an additive probability P such that $P(A) \geq P(B)$ whenever $A \geqslant B$.
- Each of the following conditions is equivalent to coherence of \geqslant .
- (c) the relations \geqslant and \geqslant^* are identical, and $\Omega > \emptyset$ ⁵
 - (d) there is a coherent lower prevision \underline{P} , defined on all gambles, such that $A \geqslant B$ if and only if $\underline{P}(A - B) \geq 0$
 - (e) there is a non-empty class \mathcal{M} of additive probabilities, such that $A \geqslant B$ if and only if $P(A) \geq P(B)$ for all $P \in \mathcal{M}$.

The equivalence of coherence and the last two conditions is a special case of the general correspondence between coherent preference orderings, coherent lower previsions and classes of linear previsions. The coherent comparative probability orderings are just those that can be generated by some coherent lower prevision \underline{P} , by $A \geqslant B$ whenever $\underline{P}(A - B) \geq 0$.⁶

4.5.3 Basic properties

A coherent comparative probability ordering \geqslant has the following properties, which hold for all subsets A, B, C, D of Ω . (These can be verified using the theorem.)

- (i) $\Omega > \emptyset$
- (ii) $\Omega \geqslant A, A \geqslant \emptyset$
- (iii) $A \geqslant A$
- (iv) $A > \emptyset$ or $A^c > \emptyset$
- (v) if $A \geqslant A^c$ then $A > \emptyset$
- (vi) if $A \supset B$ then $A \geqslant B$
- (vii) $A \geqslant B$ if and only if $B^c \geqslant A^c$
- (viii) $A \geqslant B$ if and only if $A \cap B^c \geqslant B \cap A^c$
- (ix) if $A \supset B, B \geqslant C$ and $C \supset D$ then $A \geqslant D$
- (x) if $A \geqslant B$ and $B \geqslant C$ then $A \geqslant C$
- (xi) if $A > B$ and $B \geqslant C$ then $A > C$
- (xii) if $A \geqslant B, C \geqslant D$ and $A \cap C = \emptyset$ then $A \cup C \geqslant B \cup D$; if also $A > B$ then $A \cup C > B \cup D$
- (xiii) if A, B, C, D are pairwise disjoint, $A \cup B \geqslant C \cup D$ and $A \cup C \geqslant B \cup D$, then $A \geqslant D$
- (xiv) if $A_j \geqslant B_j$ for $1 \leq j \leq n$ and A_1, \dots, A_n are pairwise disjoint then

$$\bigcup_{j=1}^n A_j \geqslant \bigcup_{j=1}^n B_j$$

- (xv) if A_1, A_2, \dots are pairwise disjoint and each $A_j \geqslant B$ then $B \approx \emptyset$.

4.5.4 Football example⁷

Because of the cancellation property (viii), it suffices in constructing a comparative probability ordering to compare disjoint events. In the football example there are three types of non-trivial comparisons between disjoint events: (1) comparison of a singleton A with its complement, which is represented in the probability simplex by the line $P(A) = \frac{1}{2}$; (2) comparison of singletons A and B , represented by the line $P(A) = P(B)$ bisecting the simplex; and (3) judgement that A and \emptyset are equally probable, represented by a side or vertex of the simplex. The natural extension $\mathcal{M}(\geqslant)$ is a convex region bounded by such lines.

The first three judgements in Example 4.1.4, $D \cup L \geqslant W$, $W \geqslant D$ and $D \geqslant L$, are comparative probability judgements. They correspond to the desirable gambles $\frac{1}{2} - W$, $W - D$ and $D - L$, which are plotted as lines in the simplex representation in Figure 4.5.4. The natural extension $\mathcal{M}(\geqslant)$ is the shaded triangle bounded by the three lines. Since this is non-empty, the partial ordering \geqslant avoids sure loss. Its natural extension \geqslant^* is the complete, coherent ordering

$$\Omega >^* W \cup D >^* W \cup L >^* D \cup L >^* W >^* D >^* L >^* \emptyset.$$

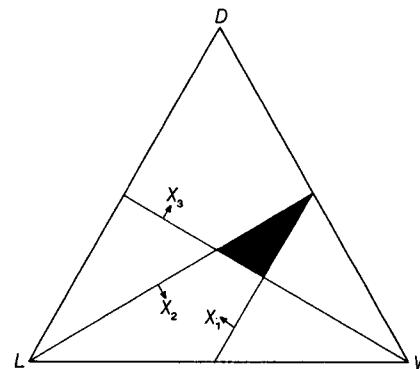


Figure 4.5.4 Model produced by comparative probability judgements

To obtain a smaller region $\mathcal{M}(\geqslant)$ it is necessary to make judgements of equality, for example that $W \approx D$ or that $L \approx \emptyset$. Larger regions $\mathcal{M}(\geqslant)$ represent incomplete orderings \geqslant^* , such as that generated by the judgements $D \cup L \geqslant W$ and $W \geqslant D$. The events W and L are incomparable under this ordering (neither $W \geqslant^* L$ nor $L \geqslant^* W$).

4.5.5 Orderings of states

Special types of comparative probability orderings can be generated by restricting the comparisons to special types of events. For example, it will often be easier to compare singleton sets (elements of Ω) than to compare sets containing many elements. Suppose that You completely order the elements of a finite space Ω , so that $\omega_1 \geqslant \dots \geqslant \omega_n$.

Any such ordering avoids sure loss. Its natural extensions are defined as follows. The linear previsions P in $\mathcal{M}(\geqslant)$ are just those satisfying $P(\omega_1) \geq P(\omega_2) \geq \dots \geq P(\omega_n)$. The extreme points of $\mathcal{M}(\geqslant)$ are the previsions P_1, \dots, P_n defined by $P_m(\omega_i) = m^{-1}$ for $1 \leq i \leq m$, $P_m(\omega_i) = 0$ for $i > m$. The lower prevision \underline{P} is therefore $\underline{P}(X) = \min\{m^{-1} \sum_{i=1}^m X(\omega_i) : 1 \leq m \leq n\}$. In the extended ordering, $A \geqslant^* B$ if and only if the i th most probable element of A is at least as probable as the i th most probable element of B , for all values of i up to the cardinality of B .

In the football example with states ordered by $W \geqslant D \geqslant L$, $\mathcal{M}(\geqslant)$ is the triangle with extreme points $(1, 0, 0)$, $(0.5, 0.5, 0)$ and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Under the extension \geqslant^* , only the events W and $D \cup L$ are incomparable.

4.5.6 Completeness

Almost all of the previous studies of comparative probability have been concerned with complete orderings.⁸ An ordering \geqslant is **complete** when, for every pair of events A and B , either $A \geqslant B$ or $B \geqslant A$. Completeness means that any two events are comparable. Any ordering \geqslant that avoids sure loss can be extended to a complete coherent ordering, and \geqslant is coherent if and only if it is the intersection of its complete coherent extensions (Theorem 3.8.5).

It can be shown that, for finite Ω , a comparative probability ordering \geqslant is coherent and complete if and only if there is an additive probability measure P that agrees with \geqslant , in the sense that $A \geqslant B$ if and only if $P(A) \geq P(B)$.⁹ Thus the lower prevision \underline{P} in Theorem 4.5.2(d) can be taken to be linear just when \geqslant is complete.

Unless Ω is a small finite space, completeness of \geqslant is a very strong requirement. You may be unable to judge which of two events is more probable. Our elicitation procedure allows You to abstain from comparing events when You find the comparison difficult, and typically Your judgements will generate only a partial ordering. But even if completeness is required, so that a coherent ordering can be represented by a linear prevision, much stronger conditions are needed to yield a unique linear prevision.

In Example 4.5.4, the coherent ordering \geqslant^* is complete, but its

corresponding lower prevision \underline{P} is quite imprecise, e.g. $\underline{P}(L) = 0$ and $\bar{P}(L) = \frac{1}{3}$. There are many additive probability measures in $\mathcal{M}(\underline{P})$ that agree with \geqslant^* .

4.5.7 Extraneous measurement probabilities

Models with greater precision can be obtained from comparative probability judgements by embedding the possibility space Ω in a larger space which contains a rich supply of ‘standard events’, with which events in Ω can be compared. This involves refining Ω (section 4.3.3). The simplest way to do this is to form a product space $\Omega \times \Omega_0$. Often Ω_0 is taken to be the unit interval $[0, 1]$, representing an **auxiliary experiment** in which a point ω_0 is selected from Ω_0 according to a uniform probability distribution, which could be physically generated by a spinning pointer on a circle.

You are then asked to compare events A (subsets of Ω) with **standard events** of the form $B = [0, x]$ in the auxiliary experiment.¹⁰ (Of course B has precise probability x .) If You are able to make these comparisons coherently for all subsets of Ω and all values of x , then Your ordering \geqslant determines a unique additive probability P on subsets of Ω . Here $P(A)$ is the unique value of x such that $A \approx [0, x]$. In principle, You can identify a precise probability model by making sufficiently many comparative probability judgements.¹¹

It may often be useful to extend the space Ω in this way, because the comparisons with standard events may enable the probabilities of events in Ω to be elicited more precisely. The attainable precision depends on the extent to which these events are comparable with the standard events.¹² In practice, comparability is limited in two ways. One limitation is that Your beliefs about Ω are often indeterminate for the good reason that You have little information about Ω , and introducing extraneous experiments will not provide more information.

The second limitation is that the standard events may be so dissimilar from the events of interest that any comparisons seem arbitrary. Would comparisons with the outcomes of spinning pointers or tossing coins really help You to assess whether the inflation rate is likely to rise in 1999? The introduction of an auxiliary experiment provides a measurement scale on which precise assessments can be expressed, but it may not help You make these assessments.

Whereas it may sometimes be reasonable to expect completeness of a comparative probability ordering on events in Ω , at least for small Ω , it is unrealistic (for the above reasons) to expect all these events to be compared with all (uncountably many) standard events. For some values of x You may be unable to judge whether A or the standard event $[0, x]$ is more probable, and then the upper and lower probabilities generated by Your

comparisons, $\bar{P}(A) = \inf\{x: [0, x] \geqslant A\}$ and $\underline{P}(A) = \sup\{x: A \geqslant [0, x]\}$, will not be equal.¹³

It is necessary, if probabilities are to be precisely measured, that the auxiliary space Ω_0 be infinite. It is more realistic to consider a finite space Ω_0 . For example, let the auxiliary experiment consist of m independent tosses of a fair coin. All events in the auxiliary space are taken to be standard events. These have precise probabilities which are multiples of 2^{-m} . In principle, You can determine the probability of any subset of Ω to within 2^{-m} by comparing it with standard events,¹⁴ although the elicited probabilities may be less precise because of incomparability. It may be possible to make all the comparisons with standard events when m is small, but that becomes more difficult as m increases and the measurement scale is refined.

4.6 Other types of assessment

It will often be convenient to elicit probabilities through judgements of classificatory or comparative probability, because these judgements can be made without a sophisticated understanding of probability theory. Their chief disadvantages are that models may be too imprecise when few judgements are made, and there may be computational difficulties, when many judgements are made, in checking that the judgements avoid sure loss and in obtaining their natural extension.

In this section we describe some other types of assessments and models that seem likely to be useful. To illustrate the models we use the football example 4.1.4, in which Ω is a three-point space, and also consider the important cases in which ω is a real-valued variable or a statistical parameter. For each kind of model, the basic issues are the same as for the general elicitation procedure (section 4.1): to explain the behavioural meaning of the judgements from which the model is constructed, to give necessary and sufficient conditions for avoiding sure loss, and to compute the natural extensions \underline{P} , \mathcal{E} and \mathcal{M} .

In the case where ω is a statistical parameter, it is important also that posterior probabilities concerning ω can be easily calculated. The most tractable models in that case are upper and lower density functions (section 4.6.3), intervals of measures (4.6.4) and classes of conjugate priors (4.6.8).

4.6.1 Assessment of probabilities

It is often possible to directly assess the upper and lower probabilities of events. These can be regarded as selling and buying prices for the corresponding gambles. Alternatively, a lower probability $\underline{P}(A)$ can be

assessed as a value of μ for which You prefer a contract worth £100 if A occurs (and nothing otherwise) to a lottery ticket which will win £100 with known chance μ . Similarly, $\bar{P}(A)$ can be assessed as a value of μ for which You prefer the lottery ticket to the contract.¹

When only finitely many upper and lower probabilities are assessed, the resulting model is finitely generated, and the method of section 4.2.1 can be used to check whether the assessments avoid sure loss and to compute their natural extension.²

Suppose, for example, that You assess upper and lower probabilities of events A_1, A_2, \dots, A_n , which form a partition of Ω .³ When Ω is finite (as in the example below), each A_j might contain a single state ω_j . When Ω is the real line, the sets A_j might be intervals.

Assuming that all the assessments $\underline{P}(A_j)$ and $\bar{P}(A_j)$ are between zero and one, they avoid sure loss if and only if they satisfy $\underline{P}(A_j) \leq \bar{P}(A_j)$ for $1 \leq j \leq n$ and $\sum_{j=1}^n \underline{P}(A_j) \leq 1 \leq \sum_{j=1}^n \bar{P}(A_j)$. The assessments are coherent if and only if they avoid sure loss and also satisfy

$$\bar{P}(A_i) + \sum_{j \neq i} \underline{P}(A_j) \leq 1 \leq \underline{P}(A_i) + \sum_{j \neq i} \bar{P}(A_j) \quad \text{for } 1 \leq i \leq n.$$

These conditions are easy to check. The natural extension of the assessments can be computed via the extreme points of \mathcal{M} , for which $P(A_j)$ is equal to either $\underline{P}(A_j)$ or $\bar{P}(A_j)$ except for (at most) one value of j . When P is used to model prior beliefs about a statistical parameter, the posterior upper and lower probabilities of events can be calculated from these extreme points, using the results of section 8.4.8.

This model can be illustrated by the football example $\Omega = \{W, D, L\}$. Consider the assessments $\underline{P}(W) = 0.27$, $\bar{P}(W) = 0.52$, $\underline{P}(D) = 0.27$, $\bar{P}(D) =$

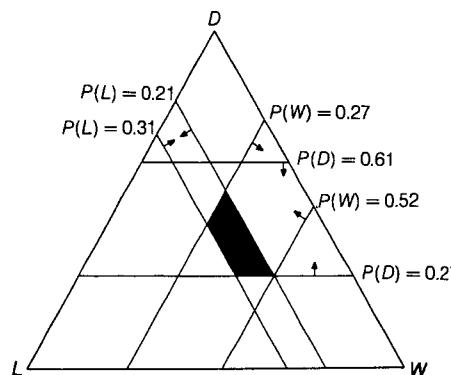


Figure 4.6.1 Model produced by assessment of upper and lower probabilities

4.6 OTHER TYPES OF ASSESSMENT

0.61 , $\bar{P}(L) = 0.21$, $\bar{P}(L) = 0.31$.⁴ In the simplex representation in Figure 4.6.1, each assessment is represented by a line parallel to one side of the simplex. The assessments do avoid sure loss, as the shaded region \mathcal{M} they determine is non-empty, but they are not coherent, as the line $P(D) = 0.61$ is not tangent to \mathcal{M} . Indeed $\bar{P}(D) + \underline{P}(W) + \underline{P}(L) = 1.09$, larger than one. In the natural extension of the assessments, $\bar{P}(D)$ is reduced to $1 - \underline{P}(W) - \underline{P}(L) = 0.52$. The line $P(D) = 0.52$ is tangent to \mathcal{M} .

The extreme points of \mathcal{M} can be easily obtained by equating $P(A_i)$ to either $\underline{P}(A_i)$ or $\bar{P}(A_i)$ for two of the three events (the third probability is then determined), and checking which of the resulting P satisfy $\underline{P} \leq P \leq \bar{P}$. This gives the four extreme points $(P(W), P(D), P(L)) = (0.52, 0.27, 0.21)$, $(0.27, 0.42, 0.31)$, $(0.42, 0.27, 0.31)$, $(0.27, 0.52, 0.21)$.

4.6.2 Assessment of probability ratios

Comparative probability judgements, that event A is at least as probable as B , can be generalized by admitting judgements of the form ' A is at least l times as probable as B ', where l is a specified positive number. The behavioural interpretation of this judgement is that the gamble $A - lB$ is almost desirable. Similarly the judgement that A is no more than u times as probable as B means that $uB - A$ is almost desirable. Since all the linear previsions P in the natural extension \mathcal{M} must satisfy $l \leq P(A)/P(B) \leq u$, l and u can be regarded as lower and upper bounds for the probability ratio $P(A)/P(B)$.⁵

A simple way of making these assessments when Ω is finite is to select some state ω_0 from Ω , and, for every other state ω , to assess upper and lower bounds $u(\omega)$ and $l(\omega)$ for the probability ratio $P(\{\omega\})/P(\{\omega_0\})$.⁶ These

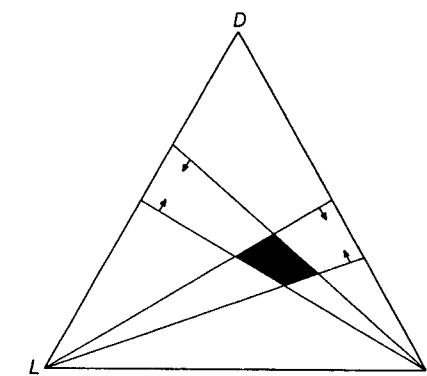


Figure 4.6.2 Model produced by assessment of upper and lower probability ratios

assessments avoid sure loss provided that $u(\omega) \geq l(\omega)$ for all states ω . If so, the extreme points of the natural extension \mathcal{M} can be easily obtained by equating the probability ratios $P(\{\omega\})/P(\{\omega_0\})$ to one of the bounds $u(\omega)$ or $l(\omega)$, for each ω . Hence we obtain

$$\underline{P}(A) = \left(1 + \sum_{\omega \in A^c} u(\omega) / \sum_{\omega \in A} l(\omega) \right)^{-1} \quad \text{and}$$

$$\bar{P}(A) = \left(1 + \sum_{\omega \in A^c} l(\omega) / \sum_{\omega \in A} u(\omega) \right)^{-1}$$

for all events A , where $l(\omega_0) = u(\omega_0) = 1$.

To illustrate this model, let $\omega_0 = D$ in the football example. Suppose You assess that W is between 1 and 2 times as probable as D , and that L is between 0.5 and 1 times as probable as D . As these assessments are implied by the judgements considered in Example 4.1.4, their natural extension \mathcal{M} contains that defined in 4.1.4. The set \mathcal{M} , which is displayed in the probability simplex in Figure 4.6.2, is bounded by four lines passing through the vertices W or L .⁷

4.6.3 Upper and lower density functions

When Ω is a continuous space, probabilities can be elicited through upper and lower bounds $u(\omega)$ and $l(\omega)$ for the ratio of probability densities $h(\omega)/h(\omega_0)$, where ω_0 is a fixed state.⁸ The functions u and l are called **upper and lower density functions**. When Ω is a subset of the real line, the upper and lower density functions can be constructed graphically, by drawing two curves such that $u(\omega_0) = l(\omega_0) = 1$.

The assessments avoid sure loss provided l has finite integral and $u(\omega) \geq l(\omega)$ for all ω . The linear previsions P in the natural extension \mathcal{M} correspond to the (unnormalized) density functions h which lie between the upper and lower densities, by $P(X) = \int_{\Omega} X(\omega) h(\omega) d\omega / \int_{\Omega} h(\omega) d\omega$. The natural extensions $\underline{P}(A)$ and $\bar{P}(A)$ can be computed by formulas analogous to those in section 4.6.2, with sums replaced by integrals.⁹

Upper and lower densities are especially useful for modelling prior beliefs about a statistical parameter ω .¹⁰ When these are combined with the likelihood function L_x determined by a statistical observation x , the model for posterior beliefs has the same form as the prior, and can be described by posterior upper and lower densities u_x and l_x where $u_x(\omega) = u(\omega)L_x(\omega)/L_x(\omega_0)$ and $l_x(\omega) = l(\omega)L_x(\omega)/L_x(\omega_0)$. Thus the prior upper and lower density functions are updated by multiplying them by the likelihood ratios $L_x(\omega)/L_x(\omega_0)$.

4.6.4 Intervals of measures

The two preceding models are special types of **intervals of measures**.¹¹ In the general model, \mathcal{M} consists of all linear previsions which correspond to (unnormalized) density functions h , with respect to a measure v that lie between specified upper and lower bounds, u and l .¹² Thus $l(\omega) \leq h(\omega) \leq u(\omega)$ for all $\omega \in \Omega$. Models 4.6.2 and 4.6.3 are special cases in which $l(\omega_0) = u(\omega_0) = 1$. The constant odds-ratio model (sections 2.9.4, 3.3.5) is another special case, in which l is a constant multiple of u .

The upper and lower probabilities of events under this model can again be computed by simple formulas,¹³

$$\underline{P}(A) = \left(1 + \int_{A^c} u(\omega) v(d\omega) / \int_A l(\omega) v(d\omega) \right)^{-1}$$

$$\text{and } \bar{P}(A) = \left(1 + \int_{A^c} l(\omega) v(d\omega) / \int_A u(\omega) v(d\omega) \right)^{-1}.$$

Consider the football example, where v is counting measure, with the assessments

$$u(W) = 6, \quad l(W) = 4, \quad u(D) = 4, \quad l(D) = 3, \quad u(L) = 3 \quad \text{and} \quad l(L) = 2.$$

These values are chosen to be consistent with the assessments in Example 4.1.4 and section 4.6.2, but they are actually more precise than those in 4.6.2, and they determine a smaller natural extension \mathcal{M} . The set \mathcal{M} is displayed in the probability simplex in Figure 4.6.4. It is bounded by six lines passing through the vertices of the simplex. These are the four lines plotted in Figure 4.6.2, together with two lines through D which represent

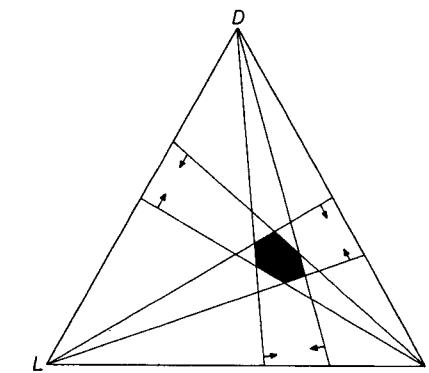


Figure 4.6.4 Model produced by assessment of upper and lower densities

the assessments

$$l(W)/u(L) = \frac{4}{3} \leq P(W)/P(L) \leq 3 = u(W)/l(L).$$

The six extreme points of \mathcal{M} can be computed by assigning either the upper or lower density to each state of Ω and normalizing to obtain a probability mass function.

Intervals of measures are attractive models for prior beliefs about a statistical parameter, because the posterior models they generate are also intervals of measures. The prior bounds l and u are simply replaced by posterior bounds $l_x(\omega) = l(\omega)L_x(\omega)$ and $u_x(\omega) = u(\omega)L_x(\omega)$, where L_x is the observed likelihood function.

4.6.5 Neighbourhood models

An approach that is popular with Bayesian sensitivity analysts is to first elicit an additive probability P_0 , and then consider possible inaccuracies in the assessment of P_0 . The second step can be formalized by constructing a neighbourhood \mathcal{M} of P_0 .

Three of the models discussed in section 2.9 correspond to different ways of defining a neighbourhood of P_0 . The linear-vacuous mixture (section 2.9.2), $\underline{P}(X) = (1 - \delta)P_0(X) + \delta \inf X$, corresponds to the δ -contamination neighbourhood $\mathcal{M}_L = \{(1 - \delta)P_0 + \delta P : P \in \mathcal{P}\}$. The pari-mutuel and constant odds-ratio models (sections 2.9.3, 2.9.4) define different types of neighbourhoods, identified in sections 3.3.5 and 3.6.3.¹⁴

The three types of neighbourhood are displayed in the probability simplex in Figure 4.6.5, for the football example with common values of $P_0 = (0.5, 0.3, 0.2)$ and $\delta = 0.2$. The linear-vacuous neighbourhood \mathcal{M}_L is a triangle bounded by the three lines $P(A) = (1 - \delta)P_0(A)$, where A is a singleton. The extreme points of \mathcal{M}_L lie on the lines joining P_0 to the vertices of the simplex.

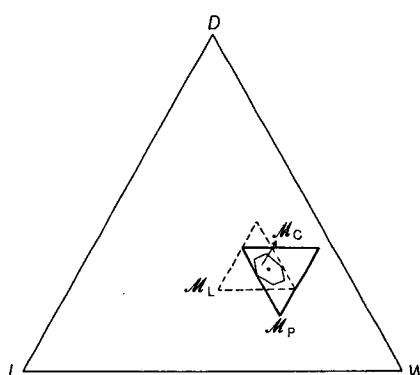


Figure 4.6.5 Neighbourhood models

4.6 OTHER TYPES OF ASSESSMENT

The pari-mutuel neighbourhood \mathcal{M}_P is a triangle obtained by reflecting \mathcal{M}_L about P_0 . It is bounded by the lines $P(A) = (1 + \delta)P_0(A)$. The constant odds-ratio neighbourhood \mathcal{M}_C is a hexagon contained in the intersection of \mathcal{M}_L and \mathcal{M}_P . It is bounded by six lines, two through each vertex, of the form $P(A)/P(B) = (1 - \delta)P_0(A)/P_0(B)$ where A and B are singletons. This is a special type of interval of measures (section 4.6.4), with $u = P_0 = (0.5, 0.3, 0.2)$ and $l = (1 - \delta)P_0 = (0.4, 0.24, 0.16)$.

Elicitation of these models involves assessment of the additive probability P_0 , selection of an appropriate type of neighbourhood, and assessment of the size of the neighbourhood (δ). Usually P_0 is elicited by standard Bayesian methods, and δ measures the probability that the elicited P_0 is incorrect, or the size of possible errors in P_0 , or the amount of information on which the model is based. In behavioural terms, the linear-vacuous model is equivalent to the judgements $\underline{P}(A) \geq (1 - \delta)P_0(A)$ for non-trivial events A , the pari-mutuel model to $\bar{P}(A) \leq (1 + \delta)P_0(A)$, and the constant odds-ratio model to the judgements that the ratio of probabilities for A to B is at least $(1 - \delta)P_0(A)/P_0(B)$.

Neighbourhood models are most appropriate when there is substantial evidence which supports an additive probability P_0 but there is incomplete confidence in P_0 . In practice it is difficult to assess additive probabilities and we prefer models which do not require them. Rather than making precise assessments to determine P_0 and adding imprecision by forming a neighbourhood, it is easier to directly elicit a model through imprecise assessments.¹⁵

Neighbourhood models have been used by sensitivity analysts to model prior beliefs about statistical parameters, in order to make Bayesian analyses more robust.¹⁶ (The constant odds-ratio prior generates a posterior of the same form, but the other neighbourhoods are somewhat more difficult to update.) However, the types of neighbourhoods that have been investigated in statistical problems appear to be less successful than other models, such as types of intervals of measures (section 4.6.4) and near-ignorance priors (section 4.6.9).¹⁷

4.6.6 Upper and lower distribution functions

Probabilities concerning a real variable X can be elicited by specifying **upper** and **lower distribution functions** (\bar{F} and \underline{F}) for X . These are just the upper and lower probabilities of the events $\{\omega : X(\omega) \leq x\}$, which we write simply as $X \leq x$, so $\bar{F}(x) = \bar{P}(X \leq x)$ and $\underline{F}(x) = \underline{P}(X \leq x)$ for all real values x . Assessments of upper and lower distribution functions can therefore be regarded as assessments of upper and lower probabilities. It may be convenient to construct these functions graphically.

Suppose, for simplicity, that the possible values of X form an interval

(a, b) . Then assessments of the functions \underline{F} and \bar{F} avoid sure loss when $\underline{F}(x) \leq 1$ and $\bar{F}(x) \geq 0$ for all real x , $\underline{F}(x) \leq 0$ for $x \leq a$, $\bar{F}(x) \geq 1$ for $x \geq b$, and $\underline{F}(y) \leq \bar{F}(z)$ for $y \leq z$. The assessments are coherent when \underline{F} and \bar{F} are non-decreasing functions which take the values zero when $x \leq a$ and one when $x \geq b$, and $\underline{F}(x) \leq \bar{F}(x)$ for all real x . More generally, the natural extension \mathcal{M} consists of all linear previsions P whose distribution function $F(x) = P(X \leq x)$ lies in the band between the upper and lower distribution functions, $\underline{F}(x) \leq F(x) \leq \bar{F}(x)$ for all real x .

For example, You might assess $\underline{F}(x_j)$ and $\bar{F}(x_j)$ at finitely many values $a < x_1 < x_2 < \dots < x_m < b$. Assuming that the assessments are non-decreasing in j and bounded by zero and one, they avoid sure loss if and only if $\underline{F}(x_j) \leq \bar{F}(x_j)$ for each j .¹⁸ The non-decreasing functions \underline{F} and \bar{F} constructed by natural extension are step functions with jumps at some of the values $a, x_1, x_2, \dots, x_m, b$. \underline{F} is right-continuous and \bar{F} is left-continuous. The natural extension \mathcal{M} contains all linear previsions whose distribution function F lies between the upper and lower assessments, $\underline{F}(x_j) \leq F(x_j) \leq \bar{F}(x_j)$ for each j .

4.6.7 Upper and lower quantiles

Instead of assessing upper and lower probabilities of pre-specified intervals, You could assess upper and lower quantiles for pre-specified probability values. The **lower ρ -quantile** of X , denoted by \underline{x}_ρ , is the infimum value of x for which $\bar{P}(X \leq x) \geq \rho$. The **upper ρ -quantile**, \bar{x}_ρ , is the supremum value of x for which $\underline{P}(X \leq x) \leq \rho$. These values satisfy $\underline{x}_\rho \leq \bar{x}_\rho$. Suppose that You make assessments of \bar{x}_ρ and \underline{x}_ρ , with the behavioural interpretation that You are willing to bet *on* the event $X \leq x$ at rate ρ whenever $\bar{x}_\rho < x$, and You are willing to bet *against* $X \leq x$ at rate ρ whenever $x < \underline{x}_\rho$. In effect, these are equivalent to assessments of upper and lower distribution functions, $\bar{F}(\underline{x}_\rho) \leq \rho \leq \underline{F}(\bar{x}_\rho)$.¹⁹

Suppose that You assess upper and lower quantiles $\bar{x}_j, \underline{x}_j$ for finitely many probabilities $0 \leq \rho_1 < \rho_2 < \dots < \rho_k \leq 1$. Assume that the possible values of X form an interval (a, b) , and that \underline{x}_j and \bar{x}_j are each non-decreasing in j . Then the assessments avoid sure loss provided $\bar{x}_1 > a$, $\underline{x}_k < b$, and $\underline{x}_j \leq \bar{x}_j$ for $1 \leq j \leq k$. The natural extensions \underline{F} and \bar{F} are step functions taking the values ρ_j , with jumps at \bar{x}_j or \underline{x}_j . Again \underline{F} is right-continuous and \bar{F} is left-continuous. The class \mathcal{M} consists of all linear previsions whose distribution function F satisfies $F(\underline{x}_j) \leq \rho_j \leq F(\bar{x}_j)$ for all j .

Suppose next that You assess a precise distribution function $F(x_j) = \rho_j$ at finitely many points, where x_j and ρ_j are each increasing in j . This is a special case of the previous model, in which the upper and lower quantiles are equal. It is also a special case of model 4.6.6, in which the upper and lower distribution functions agree at each ρ_j , and a special case of model

4.6.1, in which precise probabilities $\rho_{j+1} - \rho_j$ are assessed for each of the intervals $(x_j, x_{j+1}]$.²⁰ The natural extensions \underline{F} and \bar{F} are step functions which agree at each point x_j , but disagree at all other points in the range of X . Unless X has only finitely many possible values, the precise assessments generate an imprecise probability model.²¹

4.6.8 Classes of conjugate priors

Suppose that ω is a statistical parameter. In many statistical problems, ω parametrizes a family of density functions with a common functional form, such as binomial, Poisson or Normal densities. In these problems it is often convenient to consider a class of prior densities (with respect to Lebesgue measure) which have the same functional form, as functions of ω , as the possible likelihood functions. These are called **natural conjugate** densities.²²

Prior densities from the natural-conjugate family generate, in many common statistical problems, posterior densities that belong to the same family. A class \mathcal{M}_0 of natural-conjugate priors will then generate, by applying Bayes' rule to each density in \mathcal{M}_0 , a class \mathcal{M}_n of natural-conjugate posteriors.²³ That makes it quite easy to calculate, and draw conclusions from, the posterior class.

Suppose, for example, that ω is the chance of success in a series of Bernoulli trials. The possible likelihood functions are proportional to $\omega^m(1 - \omega)^{n-m}$. The natural-conjugate prior densities for ω are beta densities, proportional to $\omega^\alpha(1 - \omega)^\beta$ for $0 \leq \omega \leq 1$, which generate beta posterior densities. A class \mathcal{M} of beta prior densities can be simply described by a region R_0 of values for the parameters (α, β) . After observing m successes of n Bernoulli trials, You simply update the region R_0 to a posterior region R_n , by updating each point (α, β) in R_0 to $(\alpha + m, \beta + n - m)$. The inferences generated by some special kinds of models R_0 are described in sections 5.3 and 5.4.

It is difficult to elicit such a model \mathcal{M}_0 by the general method of section 4.1, because the class of desirable gambles generated by \mathcal{M}_0 is difficult to specify. But given the constraint that \mathcal{M}_0 consists only of distributions from the conjugate family, it is straightforward to elicit \mathcal{M}_0 or R_0 by the methods described in this section, e.g. through imprecise assessments of probabilities for subsets of Ω , density functions at a few values of ω , quantiles for ω (or for real-valued functions of ω), previsions and variances for ω (or bounds on other moments), or characteristics of the posterior class based on hypothetical samples, or by comparing the prior information with an 'equivalent prior sample'.²⁴

It can be dangerous to restrict attention to a family of conjugate densities merely for reasons of convenience or tractability, especially when inferences are sensitive to changes in distributional characteristics (such as the shape

of tails) that are constant within this family. It is well known that conjugate priors are typically non-robust, in the sense that they have an unduly large influence on the posterior distribution in cases of prior-data conflict, when the likelihood function is concentrated in the tail of the prior density.²⁵ This happens when the statistical observations have a Normal distribution with mean ω and known variance, so that the conjugate priors are also Normal.

These dangers are less serious, however, when classes of conjugate priors are used instead of single priors. Robustness can be achieved by choosing the class \mathcal{M}_0 to be sufficiently large, and especially to contain conjugate priors with arbitrarily long tails (e.g. Normal priors with arbitrarily large variances).

Moreover, \mathcal{M}_0 is equivalent, under our behavioural interpretation, to its convex hull, which contains finite mixtures (convex combinations) of conjugate priors. If \mathcal{M}_0 contains a reasonable variety of conjugate densities then its convex hull will contain densities with a wide variety of shapes. In fact, any prior distribution can be approximated by a finite mixture of conjugate priors.²⁶ Hence any given class of priors can be encompassed by a sufficiently large class of conjugate priors. It may sometimes be useful to replace an intractable model \mathcal{M}_0 , elicited by the other methods of this chapter, by a class of conjugate priors whose convex hull contains \mathcal{M}_0 .

4.6.9 Near-ignorance priors

In statistical problems where there is little prior information, prior probabilities should be nearly vacuous. Various constraints might be imposed on a lower prevision P to ensure that it is close to vacuous, and that the corresponding class \mathcal{M} is sufficiently large. These might require that upper and lower previsions are vacuous, $\underline{P}(X) = \inf X$ and $\bar{P}(X) = \sup X$, for various functions X of the statistical parameter or observations.²⁷ Such properties of near-ignorance can be satisfied by classes of conjugate priors, which are easy to specify and to update; see section 5.3 for examples.

4.6.10 Multivariate assessment

The models described in this section are useful primarily when Ω is a finite space or a subset of the real line. The models can be used in assessing multivariate distributions, but they will often be highly imprecise. In many problems it is essential to assess conditional probabilities, and to make structural judgements of independence, conditional independence or permutability, in order to reduce the imprecision. Various types of assessment strategy using conditional and structural judgements are described in Chapters 6 to 9.

CHAPTER 5

The importance of imprecision

It is the imprecision of our probability models that distinguishes them from standard models. Now that we have presented the basic theory of imprecise probabilities, before extending this to a theory of conditional probability and statistical inference, it is important to discuss the philosophical issues connected with imprecision. Three major issues are discussed in this chapter:

1. What are the sources of imprecision in probabilities? (sections 5.1–5.4).
2. Are precise probability models adequate, e.g. in cases of ‘complete ignorance’? Is there any justification for the Bayesian dogma of precision? (sections 5.5–5.8).
3. Are there other ways of dealing with imprecision, and how do they compare with our approach? (sections 5.9–5.13).

In section 5.1 we discuss the fundamental ideas of uncertainty, indeterminacy and imprecision. Indeterminacy is an absence of preference which can lead to indecision or arbitrary choice. Indeterminate beliefs are modelled by imprecise probabilities. Many sources of imprecision are outlined in section 5.2. Two important sources, lack of information and conflict between different types of information, are studied in more detail in sections 5.3 and 5.4, where we construct a model for learning from repeated trials. Starting from a state of near-ignorance, in which You are unwilling to bet on or against an event at any odds, You become increasingly willing to bet as You observe the occurrence of similar events. The model relates the degree of imprecision in probabilities to the amount of statistical information on which they are based (section 5.3), and to the degree of conflict between statistical and prior information (section 5.4).

In section 5.5 we examine Bayesian models for complete ignorance. Although they are called ‘noninformative’, these models have strong implications for behaviour. They are often improper and they may produce incoherent inferences and inadmissible decisions. We conclude that ignorance cannot be adequately modelled by ‘noninformative’ distributions, or by any other precise probabilities. Instead we need probability models with a high degree of imprecision.

In sections 5.6 to 5.8 we consider the arguments in favour of precision. One argument is that, in any decision problem, You must act. Imprecise probabilities may not determine an optimal action, resulting in indecision. This argument is answered in section 5.6. We consider various ways of choosing an action, including satisficing, minimax rules and arbitrary choice. In section 5.7 we survey the Bayesian axioms of precision and the arguments put forward to support them. All are completeness axioms, requiring complete comparability of events, gambles or acts. Some other, more pragmatic, arguments for precision are discussed in section 5.8, including comparisons of probability with other scientific concepts, arguments for Bayesian inference, and objections to the complexity of imprecise probabilities.

The conclusion we draw from this discussion is that, in general, precise probabilities are inadequate models for uncertainty. Some theory of imprecise probabilities is needed. Some alternatives to our theory are outlined in sections 5.9 to 5.13. Special attention is given to the interpretation of probability in these theories, and to their relationship with our approach.

Bayesian sensitivity analysis (section 5.9) is closely related to our approach because of the mathematical correspondence between lower previsions and classes of linear previsions \mathcal{M} . The linear previsions in \mathcal{M} are interpreted by sensitivity analysts as candidates for the ‘true’ or ‘ideal’ linear prevision P_T . The meaning of P_T needs to be clarified. Our fundamental objection to sensitivity analysis, and to the more elaborate theories of second-order probability and maximum entropy, is that the assumption that there are ‘ideal’ precise probabilities is both unjustified and unnecessary.

If the ideal P_T could be clearly defined, then beliefs about it could be modelled by second-order probabilities P_2 , considered in section 5.10. If it is assumed (without justification) that P_2 is also precise, then we can construct precise first-order probabilities, thus eliminating the imprecision of \mathcal{M} . Fuzzy decision analysis (section 5.11) uses membership functions to model second-order beliefs. There are difficulties in interpreting membership functions and justifying rules for combining them. It is difficult to make the extensive second-order assessments needed to apply these two theories.

Another extension of sensitivity analysis is based on the principle of maximum entropy (section 5.12), which asserts that a precise probability measure should be chosen from \mathcal{M} by maximizing entropy. This principle has some appeal for Bayesians, who must select precise prior probabilities even when there is little prior information, but in that case the precision is unwarranted.

Finally, in section 5.13 we discuss the theory of upper and lower probability developed by Dempster and Shafer. This theory, unlike the earlier approaches, is not based on sensitivity analysis. The fundamental

ideas of belief functions and multivalued mappings are compatible with our approach and can be useful in constructing coherent upper and lower probabilities. However, the theory makes extensive use of Dempster’s rule for combining belief functions. This rule does not appear to be widely applicable, and there are simple examples in which it produces a sure loss.

5.1 Uncertainty, indeterminacy and imprecision

The following sections are concerned with the sources of indeterminacy in beliefs and imprecision in probabilities. First it is necessary to define these terms and distinguish them from the more general concept of uncertainty.¹

5.1.1 Uncertainty

Uncertainty and indeterminacy are properties of beliefs, and hence they describe particular types of behavioural dispositions. We say that You are **certain** that ω_0 is the true state of affairs when You are disposed to accept every gamble X for which $X(\omega_0)$ is positive. Certainty is a rather extreme form of belief. It can be modelled by the degenerate linear prevision $P(X) = X(\omega_0)$, which assigns probability one to the state ω_0 .²

Uncertainty is just the absence of certainty. You are uncertain about Ω when, for every ω in Ω , You are not certain that ω is the true state. You should be uncertain about most things. Henceforth we will take it for granted that You are uncertain about Ω .

5.1.2 Indeterminacy

The most important distinction, for our purposes, is between two types of uncertainty. Knight (1933) calls them **determinate uncertainty** and **indeterminate uncertainty**. We will shorten these terms to **determinacy** and **indeterminacy**. Behaviourally, the distinction is simple. Say that You **prefer** one gamble X to another Y if You are disposed to choose X whenever You have a choice between them. Your beliefs about Ω are **determinate** if, for every pair of gambles X and Y defined on Ω and every positive constant δ , either You prefer $X + \delta$ to Y or You prefer $Y + \delta$ to X . This means that, for any gambles X and Y , either You have a preference for one or the other, or else the two gambles are **equivalent** for You, in the sense that You would have a preference if either gamble was improved by an arbitrarily small amount.

Your beliefs about Ω are **indeterminate** when there are gambles X and Y which are not equivalent for You, but between which You have no preference. If You are required to choose between X and Y then You will, of course,

choose one way or the other, but in cases of indeterminacy Your choice is simply not determined by Your current state of mind. Your mind is not 'made up'. Thus indeterminacy is an absence of preference which can lead, in some decision problems, to indecision.

5.1.3 Imprecision

Consider now the mathematical representation of beliefs. Uncertainty about Ω is represented by some type of probability model defined on Ω . Indeterminacy in beliefs can be represented by the various types of imprecise probability models defined in Chapters 2 and 3. **Imprecision** then refers to the mathematical properties of non-linearity of lower or upper previsions, non-additivity of lower or upper probabilities, incompleteness of a class of desirable gambles or preference ordering, and non-uniqueness of dominating linear previsions. The special case of determinacy is represented by the precise probability models that are standard in probability theory and statistics. **Precision** refers to the mathematical properties of linearity, additivity, completeness or uniqueness.

It seems clear that indeterminacy exists. A little introspection should suffice to convince You that Your beliefs about many matters are presently indeterminate. As indeterminacy appears to be a pervasive feature of real beliefs, a realistic theory of epistemic probability needs either to provide imprecise probability models through which the indeterminacy can be modelled, or to establish that it is irrational and provide methods for eliminating it. The second approach will be considered later in this chapter.

Perhaps the strongest reason for seeking precise probability models is that in practical decision problems You need to choose a unique action, which often will not be determined by imprecise probabilities. In such cases we would attempt to make probabilities more precise through a more careful and thorough analysis of the evidence. But the evidence may simply be inadequate to justify a unique decision, so that even an exhaustive analysis will result in incomplete preferences. There is then some arbitrariness in the decision, in the sense that several options are reasonable. Again, a little introspection should convince You that Your choices are often arbitrary, and that the arbitrariness is often unavoidable when You have little information on which to base the decision.³

5.1.4 Degrees of imprecision and uncertainty

Probabilities are imprecise just when upper and lower previsions disagree. A simple measure of the degree of **imprecision** concerning a gamble X is the numerical difference between the upper and lower previsions, $\Delta(X) = \bar{P}(X) - P(X)$.⁴ For an event A , the degree of imprecision $\Delta(A)$ is

bounded by the values zero, achieved just when A has a precise probability, and one, achieved just when A has vacuous probabilities $P(A) = 0$ and $\bar{P}(A) = 1$.

In cases of determinate uncertainty, the **degree of uncertainty** concerning X is often measured by the variability or dispersion of the probability distribution for X .⁵ Common measures are the **entropy** of the distribution, and the **variance** or **standard deviation** of X . The degree of uncertainty about an event A with precise probability $P(A)$ could be measured by the entropy

$$H_P(A) = -P(A)\log P(A) - (1 - P(A))\log(1 - P(A)),$$

or by the variance $V_P(A) = P(A)(1 - P(A))$. For each measure, the degree of uncertainty about A is a decreasing function of $|P(A) - \frac{1}{2}|$, maximized when $P(A) = \frac{1}{2}$. The degree of uncertainty can be measured more generally, when indeterminacy is present, by upper and lower entropies (defined in note 14 of section 5.12) or by upper and lower variances (Appendix G).

5.1.5 Uncertainty and imprecision as measures of information

An important difference between our approach and the Bayesian one is that Bayesians model a state of little information about Ω through a precise probability distribution with a high degree of uncertainty, whereas we would use an imprecise probability model with a high degree of imprecision. In the case of a single event A , Bayesians would assign precise probability $P(A) = \frac{1}{2}$ to maximize the degree of uncertainty about A , whereas we would assign vacuous probabilities $P(A) = 0$ and $\bar{P}(A) = 1$ to maximize the degree of imprecision for A . When there is no information concerning A , we would be unwilling to bet on or against A at any odds, whereas Bayesians would be willing to bet at any odds better than even.

The central issue here is whether the amount of information concerning an event is more closely related to its degree of uncertainty or its degree of imprecision. One way to answer this is to consider examples in which there is a natural measure of 'amount of information'.

Perhaps the simplest example is that in which A is a possible outcome of a binary experiment, and we observe the outcomes of n independent repetitions of the experiment. The amount of information concerning A can be simply measured by the number of observations n . Under general conditions, the degree of imprecision $\Delta_n(A)$ after obtaining n observations tends to zero as $n \rightarrow \infty$. For the specific models described in section 5.3, $\Delta_n(A)$ is actually a monotonic decreasing function of n . You will be prepared to bet more decisively on A as n increases and Your upper and lower betting rates converge. Thus the degree of imprecision does reflect the amount of information on which probabilities are based.

On the other hand, the degree of uncertainty concerning A (measured by entropy or variance) will not tend to zero as $n \rightarrow \infty$, and cannot even be expected to decrease as n increases.⁶ This suggests that amount of information is more closely related to degree of imprecision than to degree of uncertainty.

When n is large, the degree of uncertainty concerning A depends primarily on the objective chance of A occurring rather than on n . If A represents a coin landing ‘heads’ and there is a lot of information which indicates that the coin is fair, then a Bayesian will assign $P(A) = \frac{1}{2}$, yielding the same degree of uncertainty as when he has no information at all. It is a common criticism of the Bayesian approach that it fails to distinguish lack of information from knowledge of chances; both are regarded as sources of determinate uncertainty. The two sources can be distinguished when indeterminacy is admitted. Known chances are modelled by precise probabilities, while lack of information is reflected in imprecision.

5.2 Sources of imprecision

In this section we identify some features of beliefs, evidence and the assessment process which can lead to imprecision in probability models. The imprecision in a model P may reflect **indeterminacy** in actual or ideal beliefs, or it may be due to **incompleteness** of the model (section 2.10.3).¹ In the second case, P provides only partial information about beliefs that are actually (or ideally) more precise. Incompleteness is usually due to difficulties in assessment or elicitation, whereas indeterminacy reflects limitations of the available information.

Bayesian sensitivity analysts have recognized incompleteness as a source of imprecision, but they have ignored indeterminacy. In interpreting lower previsions we have been careful to admit both sources. Indeed, indeterminacy seems more important than incompleteness, from both a philosophical and a practical point of view. Incompleteness is a nuisance that we try to eliminate, as far as practicable, through careful assessment and elicitation. Indeterminacy can be eliminated, however, only by obtaining a very large amount of relevant information.

In the following subsections we discuss, informally, the main sources of imprecision in probabilities. The first five, which seem to be the most important, lead to indeterminacy, while the other nine produce incompleteness.

5.2.1 Lack of information

If there is little evidence concerning Ω then beliefs about Ω should be indeterminate, and probability models imprecise, to reflect the lack of

5.2 SOURCES OF IMPRECISION

information. We regard this as the most important source of imprecision. In the extreme case of no relevant evidence, or ‘complete ignorance’, the vacuous prevision, which is maximally imprecise, is the appropriate model. At the other extreme, when there is a large amount of relevant information, precise previsions may be reasonable. The probabilities of ‘experts’, who have extensive information about some event or variable, should be more precise than the probabilities of ignorant laymen.

Of course, this relies on some understanding of the ‘amount of information’ in a body of evidence. In statistical problems the amount of information will generally increase with sample size, so that probabilities based on large samples will tend to be more precise than those based on small samples. A specific model for Bernoulli trials is studied in detail in section 5.3.²

Often You will have substantial information about some aspects of Ω but little information about other aspects. For example, You might have extensive frequency information which specifies the occurrences of an event A in previous trials, but does not specify which of the states in A occurred. You might then assess a precise probability for A , but vacuous probabilities conditional on A .³

5.2.2 Conflicting information

Several sources of information may each provide substantial information about Ω . Typically the aggregated information will yield greater precision in probabilities than either source alone, but that may not be the case when there is conflict between the two bodies of information.

An important example involves conflict between statistical evidence and prior information about a statistical parameter. Suppose that You are initially confident that the chance of a thumbtack landing pin-up is approximately 0.5, and You then observe the outcomes of a series of tosses. If the observed relative frequency of ‘pin-ups’ is 0.2 then there is prior–data conflict. The imprecision of posterior probabilities will be greater for this sample than for a sample of the same size with relative frequency 0.5. A specific model for prior–data conflict in Bernoulli trials is examined in section 5.4.⁴

5.2.3 Conflicting beliefs

A second type of conflict occurs when conflicting probability assessments are obtained from several different sources. Conflict between expert opinions is one example of this. Suppose that two experts assess precise probabilities concerning Ω , in the form of linear previsions P_1 and P_2 . A probability model which summarizes or aggregates the expert opinions should reflect

the areas and extent of disagreement between the experts. That is achieved by the lower prevision \underline{P} which is the lower envelope of P_1 and P_2 . The imprecision of \underline{P} exactly measures the extent of disagreement between the experts on any question, since the degree of imprecision for any gamble X is $\Delta(X) = |P_1(X) - \underline{P}(X)|$, equal to the absolute difference between the experts' prices for X .⁵

Experts may disagree because they have access to different evidence or because they analyse the same evidence in different ways. More generally, the conflicting models P_1 and P_2 may be derived from two different assessment strategies applied to a common body of information.

Conflict between sources of information can sometimes be eliminated by careful analysis of the information. In the case of prior–data conflict, that might be done by re-examining the prior information and showing that prior assessments were unreasonable, or by discovering errors in the data or sampling model. In the case of several experts it might be done by questioning the judgements of one expert, or by sharing their information to encourage convergence of beliefs. In the case of several assessment strategies it might be done through more elaborate modelling which improves the original assessments. But often the conflict cannot be eliminated in these ways, and then the imprecision reflects unavoidable indeterminacy rather than incomplete modelling.

5.2.4 Information of limited relevance

Assessments of the probabilities of future events are often based on analogies between these events and past observations. The analogies are often only partial, so that the observations have limited relevance to the events of interest. For example, observations of one person tossing a thumbtack may be relevant to predicting the outcome of a toss by a different person. One simple assessment strategy is to use a Bernoulli model to derive predictive probabilities, but to increase their imprecision to reflect doubts about whether the two people have similar ways of tossing. Examples are given in sections 5.3.5 and 5.4.4.

In more complicated problems, where future events cannot be regarded as realizations of a stable stochastic process, Your current information will typically become less relevant to future events as You look further into the future, and Your probability models should become correspondingly less precise. Consider, for instance, the events A_k that the price of a commodity such as oil will increase in real terms in the year $1990 + k$. We expect the degree of imprecision concerning A_k , $\Delta(A_k) = \bar{P}(A_k) - \underline{P}(A_k)$, to increase with k , reflecting the decreasing relevance of Your current information and understanding.

5.2.5 Physical indeterminacy

You can effectively eliminate the indeterminacy in beliefs about the outcome of a future Bernoulli trial by observing sufficiently many outcomes of the Bernoulli process. However, that may not be typical of physical, social or economic processes. There appear to be limits to the precision of probability models for the future evolution of these processes, which cannot be overcome by obtaining more data. That may be because (a) the processes are physically indeterminate, governed by imprecise chances, (b) the processes are too complicated or unstable to be modelled in detail, or (c) the parameters of a complete model are not estimable from data.⁶ Such processes can be described through imprecise sampling models, which give rise to irreducible imprecision in probabilities concerning the model parameters or future observations of the process. The sources of imprecise sampling models are discussed in section 7.2 and an imprecise Bernoulli model is described in section 9.6.

We now consider some sources of incompleteness in probability models. The next four sources involve difficulties in analysing evidence and assessing probabilities, while the last four concern difficulties in elicitation.

5.2.6 Lack of introspection

Often, the more time You spend thinking about an uncertain event A , recalling and analysing relevant information, the smaller will be the degree of imprecision $\Delta(A) = \bar{P}(A) - \underline{P}(A)$. Imprecision may then result from shortage of time or from the excessive 'cost of thinking'⁷: even when a probability can be precisely assessed through a sufficiently careful analysis, the precision may not be worth the effort. In a decision problem, a little introspection may suffice to make imprecise probability assessments which determine an optimal action, and there may be no reason to seek greater precision.

5.2.7 Lack of assessment strategies

You may have plenty of relevant information but be unable to use it to construct probabilities because You lack the necessary models or assessment strategies. It is often necessary to simplify or idealize or ignore part of the evidence, or to make crude holistic judgements of probabilities. Extra imprecision should be introduced to allow for these distortions.

5.2.8 Limits to computational ability

It may be impracticable to assess ideal probabilities, even when appropriate assessment strategies are available, because of limited analytical and

computational abilities. For example, a precise probability may be determined in principle by other assessments, through natural extension. It may be computationally difficult to obtain the precise value, but easy to obtain adequate upper and lower bounds, i.e. imprecise assessments of the probability.

5.2.9 Intractable models

Your elicited probability model \underline{P} may be intractable or inconvenient, and You might prefer to adopt a less precise model \underline{Q} that is more tractable. Here \underline{Q} might be based on simplifying assumptions such as independence. In statistical problems, \underline{Q} might be chosen to be the lower envelope of a class of conjugate densities, to facilitate updating.

5.2.10 Natural extension

Previsions that are extended to a larger domain by natural extension will typically be imprecise on the larger domain, even when they are precise on the original domain.⁸

5.2.11 Choice of elicitation structure

In the elicitation procedure described in section 4.1, the elicited model \underline{P} is constructed by natural extension from a set of desirable gambles. More judgements of desirability lead to greater precision in \underline{P} , but in practice there may be limits to the number of gambles that can be considered and hence limits to the precision of \underline{P} .⁹ Precision is further limited when \underline{P} is elicited through judgements of classificatory probability (section 4.4) or comparative probability (section 4.5). When You have little information, You may be satisfied that such qualitative judgements properly reflect the indeterminacy. Inexperienced probability assessors may be unable to make sharper assessments.

5.2.12 Ambiguity

There is additional imprecision when probabilities are elicited through ambiguous judgements such as 'pretty likely' or 'about 0.2'. This kind of ambiguity is discussed in section 5.11.

5.2.13 Instability

Different elicitation methods, or the same method used at different times, may produce inconsistent probability models. This may be because

(a) underlying beliefs are unstable, (b) different information is remembered at different times, (c) the elicited probabilities are influenced by the framing or context of elicitation, or (d) the elicited probabilities are overly precise.¹⁰ In such cases we might want to model the stable aspects of beliefs and ignore the unstable ones. That can be done simply by constructing an aggregate lower prevision as the lower envelope of the elicited lower previsions.

5.2.14 Caution in elicitation

It is shown in Appendix H4 that when Your beliefs are elicited through operational measurement procedures, whereby Your reported probabilities commit You to accept particular gambles, it can be reasonable for You to report probabilities that are less precise than Your real beliefs.¹¹

5.3 Information from Bernoulli trials

In this and the following section we construct a model which relates the degree of imprecision in probabilities to:

1. the amount of information on which they are based; and
2. the degree of conflict between prior information and statistical data.

Here the statistical data consist of observations of a sequence of Bernoulli trials.

5.3.1 The beta–Bernoulli model

Suppose that You wish to assess the probability that an event will occur, and You have information concerning the past frequency of occurrence in similar situations.¹ To be specific, let A be the event that a particular thumbtack lands pin-up on the next toss. Your information is that there have been m occurrences of 'pin-up' in n previous tosses. Your upper and lower probabilities after n observations, $\bar{P}_n(A)$ and $\underline{P}_n(A)$, will obviously depend on both m and n . We have suggested that the degree of imprecision concerning A , $\Delta_n(A) = \bar{P}_n(A) - \underline{P}_n(A)$, should decrease as the amount of information, which is intuitively measured by the sample size n , increases. An exact relationship between Δ_n and n can be obtained by specifying a more detailed model.

Suppose that the past and future observations are realizations of Bernoulli trials. That is, there is a constant (unknown) chance θ of 'pin-up' on each trial, and the observations are independent (conditional on θ). The likelihood function generated by observing m successes in n trials is then proportional to $\theta^m(1-\theta)^{n-m}$.

To complete the model we need to specify prior beliefs concerning the unknown chance θ . For simplicity, these will be modelled by a class \mathcal{M}_0 of distributions from the **beta** family, which is conjugate to the Bernoulli likelihood function. The beta(s, t) distribution has probability density function $h(\theta) \propto \theta^{s-1}(1-\theta)^{t-1}$ for $0 < \theta < 1$, where the parameters (s, t) satisfy $s > 0$ and $0 < t < 1$.² The parameter t is the mean of the beta(s, t) distribution, and s determines the extent to which the mean is changed by new observations. That can be seen by computing the posterior density from the beta(s_0, t_0) prior and Bernoulli likelihood function, using Bayes' rule. The posterior after n observations is a beta(s_n, t_n) distribution with updated parameters $s_n = s_0 + n$ and $t_n = (s_0 t_0 + m)/(s_0 + n)$. Thus the posterior mean t_n is a weighted average of the prior mean t_0 and the observed relative frequency m/n , with weights proportional to s_0 and n respectively.

The class \mathcal{M}_0 of beta prior distributions for θ can be specified through a region R_0 of parameter values (s_0, t_0) . The models R_0 considered in this section have the form $R_0 = \{(s_0, t_0) : t_0 < t_0 < \bar{t}_0\}$, where s_0 is a positive constant and $0 \leq t_0 < \bar{t}_0 \leq 1$. So prior beliefs can be specified through the three parameters s_0, t_0, \bar{t}_0 .

After observing m successes in n trials, the prior region R_0 is updated to the posterior region

$$R_n = \{(s_n, t_n) : (s_0, t_0) \in R_0\} = \{(s_n, t_n) : t_n < t_n < \bar{t}_n\},$$

where $s_n = s_0 + n$, $t_n = (s_0 t_0 + m)/(s_0 + n)$ and $\bar{t}_n = (s_0 \bar{t}_0 + m)/(s_0 + n)$. The posterior region R_n has the same form as the prior region R_0 ; only the three parameters s_0, t_0 and \bar{t}_0 need to be updated.³

Now consider Your upper and lower probabilities for the occurrence of A ('pin-up') in a future trial. Because $P(A|\theta) = \theta$, Your current upper and lower probabilities for A are equal to Your current upper and lower previsions for θ . Under the above model Your prior probabilities are $\bar{P}_0(A) = \bar{t}_0$ and $\underline{P}_0(A) = t_0$, and Your posterior probabilities are $\bar{P}_n(A) = \bar{t}_n$ and $\underline{P}_n(A) = t_n$.

5.3.2 Near-ignorance priors

Any prior class \mathcal{M}_0 which generates vacuous prior probabilities for the event A , $\underline{P}_0(A) = 0$ and $\bar{P}_0(A) = 1$, will be called a **near-ignorance prior**.⁴ These vacuous probabilities reflect a complete absence of prior information concerning A . We will examine a special type of near-ignorance prior, the class of beta distributions $R_0 = \{(s_0, t_0) : 0 < t_0 < 1\}$, obtained by setting $t_0 = 0$ and $\bar{t}_0 = 1$ in the previous model. Such a class is characterized by the positive constant s_0 .

The posterior upper and lower probabilities for A generated by this near-

ignorance (s_0) prior are $\bar{P}_n(A) = (s_0 + m)/(s_0 + n)$ and $\underline{P}_n(A) = m/(s_0 + n)$. These both approach the observed relative frequency m/n as the sample size n increases. Behaviourally, this means that You are initially unwilling to bet on or against A at any odds, because of the near-ignorance property, but, as observations are made, You become willing to bet at rates approaching the observed relative frequency.

The prior degree of imprecision concerning A under the near-ignorance (s_0) model is $\Delta_0(A) = \bar{P}_0(A) - \underline{P}_0(A) = 1$. The posterior degree of imprecision is $\Delta_n(A) = \bar{P}_n(A) - \underline{P}_n(A) = s_0/(s_0 + n)$, which tends to zero as $n \rightarrow \infty$. The posterior degree of imprecision depends on the sample size n , but not on the observed relative frequency.

5.3.3 Choice of the learning parameter

The rate at which posterior imprecision decreases as sample size increases, which can be regarded as the rate at which You 'learn' from observations, depends on the parameter s_0 . We will call s_0 the **learning parameter**. Larger values of s_0 produce slower convergence of upper and lower probabilities. In fact, s_0 is just the sample size needed to reduce the degree of imprecision $\Delta(A)$ from 1 to $\frac{1}{2}$.

If the beta near-ignorance model is chosen to represent prior beliefs, the learning parameter can be assessed in various ways. One option is to assess a single posterior probability for A based on hypothetical data. For example, if the upper probability of a success in the second trial conditional on a failure in the first trial is assessed as x , then s_0 is determined by $s_0 = x/(1-x)$.

Secondly, s_0 can be assessed through prior probabilities for the joint outcomes of several future trials. If y is the prior upper probability that the first two trials have different outcomes then $y = s_0/2(s_0 + 1)$, giving $s_0 = 2y/(1-2y)$. Alternatively, s_0 can be assessed through prior beliefs concerning θ . If z is the prior upper prevision of the gamble $X(\theta) = 2\theta(1-\theta) = \frac{1}{2} - 2(\theta - \frac{1}{2})^2$ then again $z = s_0/2(s_0 + 1)$, giving $s_0 = 2z/(1-2z)$.⁵

Because each of the upper previsions x, y, z is a strictly increasing function of s_0 , smaller values of s_0 must reflect greater prior information about θ .⁶ Small values of s_0 are appropriate only when You have prior information that θ is close to zero or one; in that case the first observations are expected to be highly informative about later observations and the posteriors rapidly approach precision.⁷ When there is no reason to expect θ to be close to zero or one, larger values of s_0 are appropriate and the posteriors are relatively imprecise.

5.3.4 Laplace's rule of succession

The near-ignorance class with learning parameter $s_0 = 2$, which can be obtained from the assessments $x = \frac{2}{3}$ or $y = z = \frac{1}{3}$, is especially interesting because it contains the **uniform** prior, which was advocated by Bayes and Laplace as a model for ignorance. Laplace's famous 'rule of succession', based on the uniform prior, yields precise posterior probabilities $P_n(A) = (m+1)/(n+2)$. These are half-way between the lower and upper probabilities generated by the near-ignorance prior with $s_0 = 2$, $\underline{P}_n(A) = m/(n+2)$ and $\bar{P}_n(A) = (m+2)/(n+2)$. The behaviour of these posterior probabilities as the sample size n increases is illustrated in the table, for fixed relative frequencies m/n of $\frac{1}{2}$ and 0.

$\frac{m}{n}$	n	0	2	4	10	20	50	100	1000
$\frac{m}{n} = \frac{1}{2}$	$\bar{P}_n(A)$	1	0.75	0.67	0.58	0.55	0.52	0.51	0.501
	$\underline{P}_n(A)$	0	0.25	0.33	0.42	0.45	0.48	0.49	0.499
$\frac{m}{n} = 0$	n	0	2	4	10	20	50	100	1000
	$\bar{P}_n(A)$	1	0.5	0.33	0.17	0.09	0.04	0.02	0.002
	$\underline{P}_n(A)$	0	0	0	0	0	0	0	0

When the observed relative frequency is $\frac{1}{2}$, Laplace's rule of succession gives $P_n(A) = \frac{1}{2}$, irrespective of sample size. The posterior probability of A based on 1000 observations is the same as the prior probability, based on no information. Your rates for betting on A are exactly the same in the two cases. (This is the 'paradox of ideal evidence'.⁸)

Under the near-ignorance model, Your betting behaviour does change as sample size increases: You become willing to bet on A at odds closer to even money. The precision of Your posterior probabilities and betting dispositions directly reflects the amount of information on which they are based.

In the second example, when no successes are observed in n trials, Laplace's rule gives $P_n(A) = 1/(n+2)$. Because this probability is positive, You should always be willing to bet on success in the next trial, although only at increasingly long odds. Under the near-ignorance model the posterior lower probability of success remains zero, so that You remain unwilling to bet on success at any odds, until the first success is observed.

Laplace (1814), in a notorious application of his rule of succession, estimated that the sun had risen on 1 826 213 consecutive days since the Creation, and hence that the probability of it failing to rise on the following day was $1/1\,826\,215$. He should therefore have been willing to bet against sunrise at odds of 1 826 214 to one. We would not have been so reckless.

5.3 INFORMATION FROM BERNOULLI TRIALS

5.3.5 Discounting

Even when substantial frequency information is available, it is often of doubtful relevance to the event of interest. Suppose, for example, that A denotes the event that the home team wins a particular football match, and that You have frequency information about the results of n previous games between the two teams. You might construct a Bernoulli model but doubt its adequacy because of changes in the teams. In that case You could make use of the relative frequency $r_n = m/n$, but discount it by replacing the actual sample size n by a smaller 'effective sample size' k_n . The predictive probabilities for A are then modified to $\bar{P}_n(A) = (s_0 + k_n r_n)/(s_0 + k_n)$ and $\underline{P}_n(A) = k_n r_n/(s_0 + k_n)$, with increased degree of imprecision $\Delta_n(A) = s_0/(s_0 + k_n)$. If, for instance, there is a large amount of frequency information (so n is much larger than s_0) and we discount it by 50% (so $k_n = n/2$), then the predictive probabilities for A are modified only slightly, but the degree of imprecision is approximately doubled.

5.3.6 Informative priors

Now consider the more general class of beta (s_0, t_0) priors introduced in section 5.3.1, $R_0 = \{(s_0, t_0): t_0 < t_0 < \bar{t}_0\}$ for some positive constant s_0 , where $0 \leq t_0 < \bar{t}_0 \leq 1$. We can regard R_0 as the posterior for a near-ignorance (S) prior after observing M successes in N Bernoulli trials, where the parameters (s_0, t_0, \bar{t}_0) are related to (S, M, N) by the formulas $s_0 = S + N$, $t_0 = M/(S + N)$ and $\bar{t}_0 = (S + M)/(S + N)$.⁹

One way of assessing non-vacuous prior probabilities for A is to judge Your prior information to be equivalent to observing M successes in N Bernoulli trials. (The near-ignorance prior corresponds to $M = N = 0$, an absence of prior information about A .) These assessments are then combined with an appropriate choice of S to determine the class R_0 .¹⁰

For instance, You might judge that Your knowledge about two football teams which have never actually played each other is equivalent to observing four wins for team A in six matches. The assessments $M = 4$, $N = 6$ and $S = 2$ yield the prior class R_0 determined by $s_0 = 8$, $t_0 = 0.5$ and $\bar{t}_0 = 0.75$. The values 0.75 and 0.5 are the upper and lower probabilities that team A will win the first game between the teams.

The posterior for the informative prior, after observing m successes in n Bernoulli trials, has the same form as the prior, with M and N replaced by $M+m$ and $N+n$. If, for example, team A wins the first two games against its opponent, so $m = n = 2$, the parameters are updated to $M+m = 6$, $N+n = 8$, $s_2 = 10$, $t_2 = 0.6$, $\bar{t}_2 = 0.8$. The values 0.8 and 0.6 are the upper and lower probabilities that A will win a third game between the teams.

5.3.7 Amount of information

We have seen that, in the case of Bernoulli trials, the degree of imprecision of posterior probabilities can reflect the amount of frequency information on which they are based. This suggests that, more generally, we might be able to recover a useful measure of the ‘amount of information’ as some function of the degree of imprecision.

In the simplest case of an event A , we wish to define a measure $\iota(A)$ of the amount of information concerning A in P , in terms of the degree of imprecision $\Delta(A) = \bar{P}(A) - \underline{P}(A)$.¹¹ Since we expect imprecision to decrease as information increases, $\iota(A)$ should be a strictly decreasing function of $\Delta(A)$. Since vacuous probabilities for A reflect an absence of information, we want $\iota(A) = 0$ when $\Delta(A) = 1$, and otherwise $\iota(A) > 0$. A simple measure with these properties is defined by $\iota(A) = \Delta(A)^{-1} - 1$. This is zero just when probabilities for A are vacuous, and infinite just when probabilities for A are precise.

For the beta–Bernoulli models defined in sections 5.3.2 and 5.3.6, the measure ι has other attractive properties. Under the model 5.3.6, the prior amount of information concerning A is $\iota_0(A) = (\bar{t}_0 - \underline{t}_0)^{-1} - 1 = N/S$. (Of course this is zero for the near-ignorance priors in section 5.3.2.) After observing n Bernoulli trials, the posterior amount of information is

$$\iota_n(A) = (\bar{t}_n - \underline{t}_n)^{-1} - 1 = (s_0 + n)/s_0(\bar{t}_0 - \underline{t}_0) - 1 = (N + n)/S.$$

The difference between the posterior and prior amounts of information, $\iota_n(A) - \iota_0(A) = n/S$, can be interpreted as the amount of information about A that is provided by the Bernoulli observations. Call this the **sample information**. Thus the amounts of prior, sample and posterior information are proportional to their equivalent sample sizes N , n and $N + n$ respectively, where the proportionality constant is the reciprocal of the underlying learning parameter S .¹²

When two independent samples are combined, the two amounts of sample information are added to give the combined sample information. Two independent observations provide twice as much information as one. Moreover, by definition, sample information is added to prior amount of information to give the posterior amount of information. A near-ignorance prior provides zero information about A . All samples of the same size provide the same amount of information about A , and larger samples always provide more information.¹³

5.4 Prior–data conflict

Next we extend the beta–Bernoulli model to illustrate how conflict between prior information and statistical data can be a further source of indeterminacy.

5.4 PRIOR–DATA CONFLICT

Under the models described in this section, the posterior degree of imprecision is an increasing function of the degree of prior–data conflict. Again we will use the general beta–Bernoulli model defined in section 5.3.1.

5.4.1 Determinate prior beliefs about A

Consider first the simple case in which Your prior probabilities for A are precise. Prior beliefs about the chance of success, θ , can be represented by a class of $\text{beta}(s_0, t_0)$ distributions with fixed t_0 but varying s_0 , so $R_0 = \{(s_0, t_0) : \underline{s}_0 < s_0 < \bar{s}_0\}$, where $0 \leq \underline{s}_0 < \bar{s}_0$ and $0 < t_0 < 1$.¹⁴ Your prior probability for A is then precisely t_0 . This model is somewhat unrealistic since it implies an infinite amount of prior information concerning A , according to the measure of information defined in section 5.3.7. It will, however, help to elucidate the more realistic model that follows.

After observing n successes in n trials, the posterior region R_n consists of parameter values (s_n, t_n) with $s_n = s_0 + n$, $t_n = (s_0 t_0 + m)/(s_0 + n)$ and $\underline{s}_0 < s_0 < \bar{s}_0$. Hence we obtain posterior upper and lower probabilities $\bar{P}_n(A) = \max\{\tau_1, \tau_2\}$ and $\underline{P}_n(A) = \min\{\tau_1, \tau_2\}$, where $\tau_1 = (\underline{s}_0 t_0 + m)/(\underline{s}_0 + n)$ and $\tau_2 = (\bar{s}_0 t_0 + m)/(\bar{s}_0 + n)$, so $\tau_1 > \tau_2$ just when the observed relative frequency $r_n = m/n$ is greater than the prior mean t_0 .¹⁵

The posterior degree of imprecision concerning A can be written as

$$\Delta_n(A) = \bar{P}_n(A) - \underline{P}_n(A) = |\tau_1 - \tau_2| = |r_n - t_0|n(\bar{s}_0 - \underline{s}_0)/(n + \bar{s}_0)(n + \underline{s}_0).$$

This depends on r_n and t_0 only through $|r_n - t_0|$, which is a natural measure of the **degree of conflict** between prior information about A (represented by the precise probability t_0) and the Bernoulli observations (represented by the relative frequency of success r_n). Except in the case $r_n = t_0$, the posterior probabilities for A are imprecise. The degree of imprecision is proportional to the degree of prior–data conflict.¹⁶

5.4.2 Envelope of noninformative priors

As an example, consider the model defined by $t_0 = 0.5$, $\underline{s}_0 = 0$ and $\bar{s}_0 = 2$. This is chosen to encompass the three symmetric beta densities that are commonly used by Bayesians to model prior ignorance about θ : the uniform density ($s_0 = 2$), the improper Haldane density ($s_0 = 0$), and the intermediate density ($s_0 = 1$) advocated by Jeffreys. The posterior upper and lower probabilities for A that result from observed relative frequency $r_n = 0.2$ are shown below, for various sample sizes n .

n	0	5	10	20	50	100	1000
$\bar{P}_n(A)$	0.5	0.286	0.250	0.227	0.212	0.206	0.201
$\underline{P}_n(A)$	0.5	0.200	0.200	0.200	0.200	0.200	0.200

Because $\underline{s}_0 = 0$, the upper and lower probabilities converge at very different rates to the relative frequency 0.2. The posterior degree of imprecision $\Delta_n(A)$ is substantial for sample sizes of 10 or less, although here the imprecision arises merely from disagreement amongst Bayesians over the correct model for prior ignorance!⁴

5.4.3 A general model

We have seen that imprecision can arise from lack of information, through the models in section 5.3, or from prior–data conflict, through the model in section 5.4.1. We will now extend the previous models to allow both sources of imprecision. In the general model, prior beliefs about θ are represented by the region of beta(s_0, t_0) densities

$$R_0 = \{(s_0, t_0) : \underline{s}_0 < s_0 < \bar{s}_0, \underline{t}_0 < t_0 < \bar{t}_0\},$$

where $0 \leq s_0 < \bar{s}_0$ and $0 \leq t_0 < \bar{t}_0 \leq 1$.

Here \bar{t}_0 and \underline{t}_0 can be assessed as the prior upper and lower probabilities for A . The other parameters \bar{s}_0 and \underline{s}_0 can be assessed, as in section 5.3.3, through probabilities for simple events.⁵

The posterior upper and lower probabilities for A after observing m successes in n trials are obtained by maximizing and minimizing $(s_0 t_0 + m)/(s_0 + n)$ over (s_0, t_0) in R_0 . Define the **degree of conflict** between the prior and data to be $\kappa(r_n, \underline{t}_0, \bar{t}_0) = \inf \{|r_n - t_0| : t_0 < t_0 < \bar{t}_0\}$, which is just the distance of the observed relative frequency r_n from the interval of prior probabilities $(\underline{t}_0, \bar{t}_0)$. The posterior degree of imprecision concerning A can then be written as

$$\Delta_n(A) = (\bar{t}_0 - \underline{t}_0)\bar{s}_0/(n + \bar{s}_0) + \kappa(r_n, \underline{t}_0, \bar{t}_0)n(\bar{s}_0 - \underline{s}_0)/(n + \bar{s}_0)(n + \underline{s}_0).$$

Here $\Delta_n(A)$ is made up of two components, the first arising from lack of information and the second arising from prior–data conflict.⁶ The overall degree of imprecision is simply the sum of the degrees of imprecision from each source. The degree of imprecision depends on the observed relative frequency r_n only through the degree of conflict.

When the relative frequency falls in the interval of prior probabilities $(\underline{t}_0, \bar{t}_0)$, the degree of prior–data conflict is zero and lack of information is the only source of posterior imprecision. That is always the case for a near-ignorance prior; its prior probabilities for A are vacuous and can never be in conflict with observations. That explains why prior–data conflict does not arise for the models in section 5.3, which can all be derived from a near-ignorance prior.⁷

5.4 PRIOR–DATA CONFLICT

5.4.4 Tossing a thumbtack

To illustrate the general model, let A be the event that a thumbtack lands pin-up on a future toss. The thumbtack looks similar to others You have tested for which the chance of pin-up was around 0.5, and You assess $\underline{t}_0 = 0.4$, $\bar{t}_0 = 0.6$ and $\bar{s}_0 = 20$ because of the analogy with other thumbtacks. But it is also plausible that this thumbtack is of a different type, and You might assess $\underline{s}_0 = 5$ to allow for this possibility. The difference between \underline{s}_0 and \bar{s}_0 reflects Your doubts about how seriously to take the analogy, and about how much weight to give to the prior information about other thumbtacks, relative to the Bernoulli observations, in determining posterior probabilities. The assessment of \underline{s}_0 will have no effect on posterior probabilities if the analogy is correct and observations are consistent with prior beliefs, but the large difference between \underline{s}_0 and \bar{s}_0 will be reflected in greater posterior imprecision if the analogy is mistaken and there is prior–data conflict. That is illustrated by the following table of posterior probabilities and degree of imprecision for A , after observing relative frequencies $r_n = 0.4, 0.2$ or 0 , for various sample sizes n .

n	0	5	10	20	50	100	1000	
$r_n = 0.4$	$\bar{P}_n(A)$	0.6	0.56	0.53	0.50	0.46	0.43	0.404
	$\underline{P}_n(A)$	0.4	0.40	0.40	0.40	0.40	0.40	0.400
	$\Delta_n(A)$	0.2	0.16	0.13	0.10	0.06	0.03	0.004
$r_n = 0.2$	$\bar{P}_n(A)$	0.6	0.52	0.47	0.40	0.31	0.27	0.208
	$\underline{P}_n(A)$	0.4	0.30	0.27	0.24	0.22	0.21	0.201
	$\Delta_n(A)$	0.2	0.22	0.20	0.16	0.10	0.06	0.007
$r_n = 0$	$\bar{P}_n(A)$	0.6	0.48	0.40	0.30	0.17	0.10	0.012
	$\underline{P}_n(A)$	0.4	0.20	0.13	0.08	0.04	0.02	0.002
	$\Delta_n(A)$	0.2	0.28	0.27	0.22	0.13	0.08	0.010

The posterior upper and lower probabilities both approach r_n as n increases, as with the near-ignorance models in section 5.3, but the degree of imprecision now depends on the degree of prior–data conflict as well as on sample size. For fixed sample size, the degree of imprecision is minimized when there is no prior–data conflict ($r_n = 0.4$, the first case in the table). In that case, the degree of imprecision decreases as sample size increases.

The degree of imprecision is maximized when prior–data conflict is maximized ($r_n = 0$, the last case in the table). Then the degree of imprecision

from a sample of 5 is substantially larger than the prior degree of imprecision, and it requires a sample of 25 to provide enough information to compensate for the effect of prior–data conflict. The imprecision when $r_n = 0$ is roughly twice as large as when $r_n = 0.4$, indicating that roughly the same amount of imprecision arises from prior–data conflict as from lack of information.⁸

5.4.5 *The practical effect of conflict*

Under the general model of section 5.4.3, prior–data conflict produces greater imprecision in posterior probabilities. A high degree of conflict will often have the practical effect of forcing You to re-examine the prior information and statistical data, because it leads to indeterminate conclusions or indecision. That effect seems both realistic and useful. Conflict between different sources of information usually calls for careful analysis rather than some automatic resolution.⁹

Precise prior probabilities do provide an automatic resolution of conflict, by assigning precise weights to the two sources of information. Precise priors generate precise posteriors and unique decisions; no further analysis of the conflict is required. On the other hand, imprecise probability models allow You to suspend judgement over what weights to give to the two sources of information. It may often be sensible to suspend judgement in this way, because the extra judgements will not be needed when the two sources of information are consistent. Further analysis is needed just when the conflict is sufficiently large to produce indeterminacy and indecision.¹⁰

5.5 Bayesian noninformative priors

We have suggested that probabilities should be imprecise when they are based on little information. The extreme case of ‘complete ignorance’, meaning a complete absence of relevant information concerning the possibility space Ω , should be modelled by vacuous probabilities, which are maximally imprecise. Contrary to this, some Bayesians have suggested that complete ignorance should be represented by certain precise probability distributions called **noninformative** distributions.¹ These are distributions with (in some sense) a maximal degree of uncertainty. In this section we consider whether noninformative distributions are adequate models when there is little or no relevant information.

Are You ever completely ignorant about Ω ? It is arguable that, whenever You understand the meaning of the possible states ω in the space Ω , You must have some relevant information that can be used to assess probabilities. We will argue in section 7.4.1 that, in typical statistical problems, You cannot learn from data about a statistical parameter ω if You are initially completely

ignorant about its value. Since learning does seem to be possible in such cases, we infer that ‘complete ignorance’ about statistical parameters is rare.²

However, there do seem to be genuine cases of complete ignorance, such as those in which You have never contemplated the states in Ω , or You do not understand their meaning. (Are more eggs gronks than zarks?)

In this section we will assume that You are completely ignorant about Ω , and compare Bayesian noninformative distributions with vacuous probability models. Even if it is denied that the extreme case of complete ignorance is ever realized in practice, this comparison is illuminating. There are surely many problems in which You have little useful information. Then Bayesians would adopt precise probability models with a high degree of uncertainty, which are close to noninformative distributions, whereas we would adopt highly imprecise probabilities, which are close to vacuous. Many of our criticisms of noninformative distributions apply also to other Bayesian models.³

5.5.1 *Ignorance about finite spaces*

First suppose that the space $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ is finite. In this case, Bayesians are agreed that the appropriate noninformative distribution is the **uniform** distribution, which assigns the same precise probability $P(\{\omega\}) = n^{-1}$ to each state. The uniform distribution has traditionally been defended through principles of indifference or insufficient reason.⁴ A weak version of these principles is contained in the following.

Symmetry principle: If You are completely ignorant concerning Ω then You have no information which favours any possible state over any other, and therefore Your probability model should be symmetric in the states.

In this form, the principle seems sound. If we require also that the probability model be precise, this leads to the uniform distribution as the unique representation of complete ignorance.

However, as was well known in the nineteenth century, precision is incompatible with more compelling requirements. For instance, an event A can be expressed as a set of possible states ω in many different ways, depending on how much detail is included in ω . Since the meaning of the event is invariant under refinements of the states ω , its probability should also be invariant under refinements. This leads to the following.

Embedding principle: The probability assigned to an event A should not depend on the possibility space Ω in which A is embedded.⁵

For example, if A is the event that the first ball drawn from an urn is blue, where the composition of the urn is completely unknown, then Your probability for A should not depend on how You enumerate the alternatives, especially as there is no natural way of doing so.⁶

The symmetry and embedding principles are not compatible with precision of probabilities, but they are compatible with coherence. In fact, the three principles (symmetry, embedding and coherence) are satisfied by a unique probability model.⁷ This is the **vacuous** model, defined by $P(A) = 0$ and $P(A) = 1$ for all non-trivial events A , which seems to have all the properties appropriate in a model for complete ignorance.⁸ It has minimal precision amongst all coherent models, and it is therefore a natural model for minimal information.

Another fundamental criticism of the uniform distribution is that it gives rise to sharp betting rates and complete preferences and therefore seems highly ‘informative’. Vacuous probabilities, on the other hand, say nothing at all about Your behavioural dispositions; they are truly ‘noninformative’.

5.5.2 Ignorance about a chance

Noninformative distributions are used most frequently by Bayesians as prior distributions for a statistical parameter θ . In most statistical problems the parameter space Θ is infinite, so we now consider noninformative priors on infinite parameter spaces. Consider first the case where θ is an unknown chance, such as the chance of a thumbtack landing pin-up, so that Θ is the unit interval $[0, 1]$.

In this case there is disagreement amongst Bayesians concerning the appropriate noninformative prior. At least five different priors have been advocated.⁹ One is the uniform prior, with constant density function $h_1(\theta) = 1$ on the unit interval, advocated by Laplace (1812) and more tentatively by Bayes (1763). Jeffreys (1983) and Perks (1947) suggested the density function $h_2(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$, which has been supported by Box and Tiao (1973) and Bernardo (1979) for statistical problems where θ is the chance of success in a fixed number of Bernoulli trials. For problems in which Bernoulli trials are continued until a fixed number of successes are obtained, Bernardo, Box and Tiao advocate a different density function, $h_3(\theta) \propto \theta^{-1}(1 - \theta)^{-1/2}$. This is ‘improper’ in the sense that it is not integrable over Θ . Another improper prior, $h_4(\theta) \propto \theta^{-1}(1 - \theta)^{-1}$, is defended by Haldane (1945), Novick and Hall (1965), Jaynes (1968) and Villegas (1977).¹⁰ Finally, Zellner (1977) has proposed the proper density $h_5(\theta) \propto \theta^{\theta}(1 - \theta)^{1-\theta}$.

These five densities are very different, and they can generate quite different inferences from a small sample of observations.¹¹ Each of the densities can be defended by showing that it alone has certain properties that one would

expect of a ‘noninformative’ prior, such as invariance under some kinds of transformations or optimization of some measure of information. It follows that each of the densities lacks some desirable properties, and this casts doubt on each one’s adequacy as a model for ignorance. It is difficult to accept that any of these densities is the ‘correct’ model for ignorance about the chance θ .

The proposals of Jeffreys (1946), Box and Tiao (1973), Zellner (1977) and Bernardo (1979) are especially disturbing because they violate the discrete likelihood principle (8.6.1).¹² Suppose that θ is the chance of success in Bernoulli trials. You can choose to observe the number of successes (M) in n trials, or else the number of trials (N) needed to give m successes. According to Bernardo, Box and Tiao, Your prior distribution for θ , when You have no prior information, should depend on which of the two experiments you intend to carry out. If, for example, You choose to make two observations and observe one success ($n = 2, M = 1$) then You should use the prior density h_2 and obtain posterior density proportional to $\theta^{1/2}(1 - \theta)^{1/2}$. But if You choose to sample until You obtain the first success, and this requires two trials ($m = 1, N = 2$), then You should use prior density h_3 and obtain posterior density proportional to $(1 - \theta)^{1/2}$. The likelihood function is the same in both cases, but Your conclusions may be quite different. (For instance, the posterior probability that $\theta \geq \frac{1}{2}$ is 0.5 in the first case but 0.35 in the second.)

5.5.3 Ignorance about a location parameter

Consider next the important problem in which θ is a real **location parameter**, which indexes probability density functions of the form $f(x|\theta) = f_0(x - \theta)$. Here the improper uniform density $h(\theta) = 1$ is widely used by Bayesians as a noninformative prior.

For example, the observation x might have a Normal distribution with mean θ and variance 1, so that f_0 is the standard Normal density. The uniform prior then generates a Normal posterior distribution for θ with mean x and variance 1.

The uniform prior is often defended as follows.¹³ Suppose that θ is a location parameter for the real observation x . By translating the scale on which the measurement x is made, we could equivalently observe $y = x + c$. If we also translate θ to $\psi = \theta + c$ then ψ is a location parameter for y , with f_0 unchanged, and the new problem involving y and ψ is identical in structure to the original problem involving x and θ . Assuming that You are completely ignorant about the location parameter θ , You should be equally ignorant about ψ . If complete ignorance is represented by a noninformative prior probability P_0 , the probability $P_0(A)$ that ψ falls in a set A should be the

same as the probability that θ falls in A , which happens just when ψ falls in $A + c = \{a + c : a \in A\}$. Thus we require that a noninformative prior probability P_0 be **translation-invariant**, $P_0(A) = P_0(A + c)$ for all real numbers c and measurable sets of reals A . This does seem to be a reasonable requirement.

The second condition imposed by Bayesians is that a noninformative prior P_0 should be a countably additive measure. The two requirements of translation-invariance and countable additivity imply that P_0 is a multiple of Lebesgue measure.¹⁴ But Lebesgue measure cannot be normalized to obtain a proper probability measure, so the two requirements cannot be satisfied by any probability measure. Nevertheless, Bayesians do use Lebesgue measure, and its uniform density, as a ‘noninformative prior’. It is ironic that the search for noninformative prior probabilities has led Bayesians to adopt ‘priors’ that are not probabilities.¹⁵

5.5.4 Objections to the improper uniform prior

(a) Behavioural meaning

Because the uniform prior P_0 is improper and incoherent, it cannot be given the usual behavioural interpretation of probabilities. However, Hartigan (1983) has extended the behavioural interpretation to improper priors like P_0 , by taking a gamble X to be desirable just when $P_0(X) \geq 0$. Applying this to the gamble $X = \mu - A$, where μ is positive and A is a bounded set of reals, we find that $P_0(X)$ is infinite. This presumably implies that You should be willing to sell the gamble A for any positive price μ , no matter how small. When You adopt the ‘noninformative’ uniform prior You must be almost certain, on this interpretation, that the absolute value of θ is greater than 10^{10} , in the sense that You are willing to bet on this event at any odds. Equivalently, You are willing to bet at any odds that θ^{-1} lies in an arbitrarily small interval around zero. That seems absurd.¹⁶ Such confident betting behaviour reveals substantial information, rather than complete ignorance, about θ .

(b) Posterior precision

Bayesians may try to avoid the embarrassingly strong behavioural implications of P_0 by declining to give it any direct behavioural interpretation, and using it merely as a device for generating a posterior density through Bayes’ rule. (The posterior density can be given the usual behavioural interpretation, provided it is a proper density.) There are several objections to this approach:

1. Why should the posterior distribution generated by P_0 be used to guide behaviour when P_0 itself is not?

2. Why should prior ignorance about θ , plus a single observation from the model parametrized by θ , generate a precise (highly informative) posterior distribution for θ ? This happens only because the uniform prior for θ implicitly contains very substantial information about θ .
3. One can still ask what prior beliefs and behaviour are appropriate in a state of complete ignorance, and how these are related to the uniform prior. In fact, the sampling models and posterior probabilities provide information, summarized in their natural extension, about prior beliefs. It turns out that the strong betting dispositions that were criticized in (a) can be recovered from the sampling models and posteriors by natural extension; see Example 8.2.9.

(c) Incoherence

Inferences from the uniform prior incur sure loss. That is, whenever θ is a location parameter, the posterior densities generated by P_0 can be combined with the sampling models to produce a sure loss. (See section 7.4 for a proof, examples and discussion of the incoherence.) This result does not rely on any kind of behavioural interpretation of P_0 . The sure loss is generated by the posteriors and sampling models without reference to the prior.¹⁷

(d) Inadmissibility

In statistical decision problems, the uniform prior can generate poor decision rules. Suppose, for example, that the observation x has Normal distribution with mean θ and variance 1, and You wish to estimate θ^2 . Assume that Your loss is measured by the square of the estimation error. Then the Bayes estimator generated by P_0 , which is $x^2 + 1$, has uniformly higher risk function (for all values of θ) than the estimator $x^2 - 1$. Here P_0 gives rise to a decision rule that is ‘uniformly inadmissible’. That is not surprising, in view of the incoherence noted in (c).¹⁸

(e) Invariance

The uniform prior P_0 is invariant under linear transformations of θ , but not under other one-to-one transformations. Real-valued functions such as θ^3 , θ^{-1} or $\sinh(\theta)$ do not have uniform densities. If You are completely ignorant about the value of θ , however, then You seem to be equally ignorant about the values of these functions. Information about θ^3 seems to be equivalent to information about θ .¹⁹

(f) Dependence on sampling model

Advocates of noninformative priors respond to this lack of invariance by arguing that the appropriate noninformative prior for θ must depend not only on the mathematical form of the parameter space Θ , but also on the

role of θ in indexing the sampling densities $f(x|\theta)$. The uniform prior is appropriate for a location parameter θ , but not for θ^3 , θ^{-1} or $\sinh(\theta)$ since these are not location parameters. Most authors follow Jeffreys (1946) in taking the noninformative density to be proportional to the square root of Fisher's expected information about θ , which depends on the sampling model. This definition is invariant under one-to-one differentiable transformations of θ .²⁰

There are strong objections to the dependence of noninformative densities on sampling models:

1. Why should the model for ignorance about θ depend on what statistical experiment (if any) will eventually be carried out to provide information about θ ? Many different experiments may be feasible. When θ is the chance of success in Bernoulli trials, for example, the noninformative prior for θ will depend on what stopping rule is used to terminate the trials.²¹
2. The discrete likelihood principle can be violated, as seen in section 5.5.2.
3. This approach applies only to statistical problems. What if You are completely ignorant about a quantity which is not a statistical parameter?

(g) Several parameters

As the dimension of the parameter space increases it becomes more difficult to choose a noninformative prior and to have any confidence that it will produce reliable inferences. There are difficulties even in the apparently simple case of a location-scale family, where one parameter μ is a location parameter and another σ is a scale parameter.²² The noninformative density obtained by Jeffreys' method or by an invariance argument is $h_1(\mu, \sigma) \propto \sigma^{-2}$, but a more popular choice is $h_2(\mu, \sigma) \propto \sigma^{-1}$, the product of the uniform density for μ and the standard noninformative density for σ . The second prior, which is widely used, generates incoherent inferences in common statistical problems. (See examples 7.4.6 and 7.5.8.)

(h) Improper posteriors

Improper priors may produce improper posteriors. An important example is the Normal hierarchical model, where observations x_1, x_2, \dots, x_n are generated independently from Normal distributions with means θ_j and variance one, and the means θ_j are independently generated from a Normal distribution with unknown mean μ and variance σ^2 . It is natural to adopt the standard noninformative prior for location-scale families, $h_2(\mu, \sigma) \propto \sigma^{-1}$, but that generates a posterior density for (μ, σ) that is improper, whatever the observations.²³

(i) Approximations to proper priors

Some authors seek to avoid the problems arising from the use of improper priors by regarding them merely as convenient approximations to actual priors which are proper and coherent.²⁴ The idea seems to be that, when there is little prior information about θ , the actual prior for θ should have high degree of uncertainty. For instance, it might be a Normal prior with large variance. It might then be convenient to approximate this by an improper uniform prior. The earlier arguments indicate, however, that the little prior information about θ cannot be adequately modelled by a precise probability distribution with a high degree of uncertainty.

(j) Hypothesis testing

To illustrate the previous point, consider the problem of testing a point null hypothesis concerning a real parameter θ . For simplicity, suppose the observation x has Normal distribution with mean θ and variance 1. You are interested in the hypothesis H_0 that θ is zero, and You assign positive prior probability ρ_0 to H_0 . Let H_1 denote the alternative hypothesis, that θ is non-zero. If You have no information about the possible alternatives then You might assess a prior distribution for θ conditional on H_1 that has high degree of uncertainty, such as a Normal distribution with mean zero and very large variance η^2 . But then it can be verified that the posterior probability of H_1 is less than $\eta^{-1}(\rho_0^{-1} - 1)\exp(x^2/2)$. As η increases, presumably modelling a decreasing amount of prior information about θ , the posterior probability of H_1 tends to zero. The inferences from the improper uniform prior are obtained in the limit as $\eta \rightarrow \infty$, and then the posterior probability of H_1 is zero whatever the observation x !

These inferences are unacceptable. Even when there is no prior information about alternatives to H_0 , You would wish to assign high posterior upper probability to H_1 after observing $x = 5$. The problem is that the high-variance and uniform priors are not modelling ignorance about θ . Rather, they are expressing a strong prior belief that $|\theta|$ is very large.²⁵

(k) Reference priors

Box and Tiao (1973) and Bernardo (1979) have argued that a noninformative prior should be regarded as a reference prior, that is, 'a prior which it is convenient to use as a standard'²⁶ in analysing statistical data. This raises the question of why the uniform prior (or any other noninformative prior) should be singled out as a 'standard'; the preceding list of defects suggests that it is an especially poor standard.

Box, Tiao and Bernardo argue, in fact, that noninformative priors are suitable reference standards because they produce 'reference posterior distributions which approximately describe the kind of inferences which one

is entitled to make with little relevant initial information'.²⁷ These authors apparently fall back on the assumption criticized earlier, that 'little initial information' should be modelled by a noninformative prior, at least as a good approximation to some proper prior with a high degree of uncertainty.

(1) *Sufficiently large samples*

The last line of defence for the uniform prior is the argument that, when the statistical data are sufficiently informative so that the likelihood function is sharply peaked, it really 'doesn't matter' what prior is used, because all 'reasonably smooth' prior densities will generate approximately the same posterior density.²⁸ It then suffices to choose any smooth prior density, but the uniform density will often be the most convenient choice to simplify calculation of the posterior.

This argument supports the uniform prior only in those cases where it produces approximately the same conclusions as the highly imprecise prior constructed from a sufficiently large class of prior densities. The uniform prior may indeed lead to reasonable inferences when the data are highly informative, because in that case the posterior probabilities generated by the imprecise prior are nearly precise, and can be well approximated by the precise posterior from the uniform prior. When the data are not highly informative, the inferences from the uniform prior are unreasonably precise.

5.5.5 *The nonsense of noninformative priors*

Our conclusion is that the quest for Bayesian noninformative priors is futile. So-called 'noninformative priors' are not 'noninformative' (they have strong implications for behaviour), and are often not even 'prior probabilities' (when they are improper). The problem is not that Bayesians have yet to discover the 'truly' noninformative priors, but rather that no precise probability distribution can adequately represent ignorance. That is especially clear when it is recognized that complete ignorance can be properly modelled by the vacuous probabilities, and near-ignorance by near-vacuous probabilities, which meet the objections to noninformative priors.

One might wonder why noninformative priors are still used and defended, in spite of their many serious defects. We conjecture that their persistence is due to some combination of the following:

1. The problem of little or no prior information is important in theory and common in practice.
2. According to the Bayesian dogma of precision, any state of uncertainty, even complete ignorance, can be represented by some precise probability distribution.

3. A noninformative prior has at least some desirable properties (such as invariance) that any other precise prior must lack.
4. There is a need for simple, automatic statistical methods that do not require assessments of prior information from the user.
5. Some Bayesians wish to develop objective statistical methods, which require 'objective' or 'logical' prior probabilities.
6. In some important problems, inferences based on noninformative priors are formally identical to frequentist inferences such as those obtained in the Neyman–Pearson theory of confidence intervals (section 7.5). This gave the impression that Bayesians could reproduce the 'successes' of frequentist statistics, and seemed to confirm that noninformative priors gave reasonable answers.²⁹
7. Adopting a uniform prior density allows Bayesians to interpret the normalized likelihood function as a posterior density, which makes computations simple.
8. When the data are highly informative, a noninformative prior can give reasonable answers.

If we discard noninformative priors, how can we model prior beliefs when we have little information about a statistical parameter? Vacuous prior probabilities will not be useful in that case because they generate vacuous posterior probabilities (section 7.4.1). But imprecise priors which are close to vacuous, such as the near-ignorance priors defined in section 5.3, can generate realistic posterior probabilities whose degree of precision reflects the amount of information provided by the data. The essential property of such models is that they are highly imprecise and therefore genuinely 'noninformative'. Near-ignorance models promise to achieve some of the objectives of noninformative priors while avoiding most of their defects. It will, of course, require substantial effort to develop and choose between the possible models for near-ignorance.³⁰ A final criticism of the theory of noninformative priors is that it has deflected attention from this task.

5.6 Indecision

In the next three sections we examine the arguments that have been put forward to support the Bayesian dogma of precision, that probability models should always be precise. The first argument to be considered is that precise probabilities are needed in decision problems in order to determine an optimal action.¹ This argument seems to underlie many of the Bayesian theories of probability that are outlined in section 5.7.

Consider a decision problem with a specified set of feasible actions. We say that You are **decisive** in this problem when there is a 'best' action that

You prefer to all the other feasible actions. Suppose that You analyse the decision problem by defining an appropriate possibility space Ω and assessing probabilities and utilities associated with Ω . If Your probabilities and utilities are precise then You can, in most cases, determine a best action by maximizing expected utility. But if either Your probabilities or Your utilities are imprecise then Your preferences may be incomplete, and You may be indecisive. In that case, Your analysis fails to determine a best action, although it may be helpful in identifying some preferences and ruling out some unreasonable actions.

5.6.1 Decision versus precision

The connection between precise probability models and decisive action is not as close as it may appear. Precision of probabilities is neither necessary nor sufficient to determine an optimal action, even supposing that the space Ω is given and the associated utilities are precise. It is not sufficient because there may not be a unique action that maximizes expected utility. When the set of feasible actions is infinite, for example, there may be no action that achieves the supremum expected utility.

As another example, suppose that You must identify which of two states of affairs, ω_1 or ω_2 , is the true state. You will receive one unit of utility if You correctly identify the state, zero otherwise. If there is no information about the state, a Bayesian might adopt the uniform distribution $P(\{\omega_1\}) = P(\{\omega_2\}) = \frac{1}{2}$. He would then find that the two feasible actions have the same expected utility, so he may be indecisive between them.²

More importantly, precise probabilities are not necessary for decisive action in a specific decision problem. In the last example, the assessments $P(\{\omega_1\}) = 0.6$ and $\bar{P}(\{\omega_1\}) = 1$ are highly imprecise but decisive, as they lead to a preference for option ω_1 .³ It might be difficult here to assess a precise probability for ω_1 , and it is unnecessary to do so. In many decision problems it will be much easier to determine a best action, through imprecise assessments of probabilities and utilities, than to determine precise probabilities.

5.6.2 Choice versus preference

It is essential to distinguish between choice and preference.⁴ A choice is a decision about how to act, that is made in a specific context (time, place and set of options). A preference is an underlying disposition to choose in a particular way. Actions are directly observable. Choices are ‘observable’ on the assumption that nothing intervenes between choice and action. Preference is not observable; it is a theoretical concept, like beliefs and

values, to which it is related. If You prefer action a_1 to a_2 then You will choose a_1 rather than a_2 . On the other hand, You can choose a_1 over a_2 without having any preference between them. A choice can be arbitrary, in the sense that it is not determined by Your preferences, beliefs and values.

It is clear that arbitrary choices are made in everyday life: choosing what shirt to wear or what cereal to eat for breakfast. Even important choices, concerning such matters as employment, marriage or investment, need not be determined by preferences. That is indicated by psychological feelings of indecision, instability ('changing one's mind') and sensitivity to 'framing' (how the decision problem is analysed or presented).

5.6.3 Personalist Bayesians

To understand how Bayesians obtain unique decisions, we need to distinguish between several Bayesian theories. The basic distinction is between the personalist and logical interpretations of probability. One personalist view is that, when any decision problem arises, You already have (in Your mind) beliefs, values or preferences which would justify a unique decision, and these merely need to be discovered through careful elicitation. That is, You merely need to discover Your ‘true preferences’.⁵ But it is hardly credible that we typically have complete preferences. Rather, these need to be constructed by analysing the evidence.

A second view, probably held by most personalists, is that You should choose a best action by making precise, personal assessments of probabilities and utilities, which must be coherent but are otherwise arbitrary, and maximizing expected utility. But this simply shifts the arbitrariness in making decisions from the choice of action to the choice of probabilities and utilities. To remove the arbitrariness, personalists need to show that precise probabilities are determined in some way by the available information. (If they could do so, they would no longer be personalists.)

This personalist approach resembles ours in that both admit some arbitrariness in choice. If the sole purpose of an analysis is to choose an action, then it may be quite reasonable to do so in the personalist way, by choosing precise probabilities and utilities. (Although, as noted above, this will usually require far more effort than is needed to choose an action.) But even when choosing an action is Your primary purpose, You might also be interested in obtaining insight into how probabilities and preferences are influenced by the various sources of information, and where further information is most needed to reduce the arbitrariness in choice. This can be done only by explicitly modelling the arbitrariness, through imprecise probabilities which reflect the limitations of the information on which they are based.⁶ This becomes more important as the purposes of analysis move

towards inference rather than decision. The precision of Bayesian inferences can be seriously misleading when there is little information.

5.6.4 Logical Bayesians

According to logical Bayesians like Carnap and Jeffreys, there is a unique probability measure that properly represents the available information concerning Ω . Because these probabilities are supposed to be precise, they will often determine a uniquely rational decision.

There are three main objections to this approach. First, it is assumed that utilities are also precise. But there may be the same kind of arbitrariness in the choice of these (presumably personal) utilities, and hence in decisions, as for personalist Bayesians. Second, logical probability measures have been defined only for very simple kinds of information. Third, the measures seem to be inadequate even in those cases where they have been defined. In the most important case of complete ignorance, which was discussed at length in section 5.5, it seems that no precise probability measure is adequate.⁷ Why should we expect, when we have no information about a problem, to be able to rule out all but one of the feasible actions as irrational?

It is instructive that in three of the most serious attempts to develop a logical interpretation of probability, by Keynes (1921), Kyburg (1974a) and Levi (1980), the probability models are typically imprecise. These theories thereby admit some degree of indecision and arbitrary choice.

5.6.5 Bayesian sensitivity analysis

Those Bayesians who advocate sensitivity analysis implicitly recognize that You cannot always be decisive. If You assess a range of precise probabilities, and different probabilities in this range support different actions, then Your analysis fails to determine a best action. Some Bayesians might insist that in such cases You could always narrow the range of probabilities to obtain a unique action, but this may require the kind of arbitrary choice that the sensitivity analysis was designed to avoid. Sensitivity analysts must accept the same kind of incomplete preferences and indecision as we do.⁸

5.6.6 How to choose?

In making decisions, it is often useful to construct preferences between options by assessing probabilities and utilities. Typically, the assessments will be imprecise, the preferences will be incomplete, and there will be more than one reasonable (maximal) option. There will be some arbitrariness in Your choice from amongst the maximal options. It is natural to seek further

rules or guidance which might reduce the arbitrariness and perhaps determine a unique decision.

There are two types of strategy which can be used to make a decision: (i) do more analysis in order to extend Your preference ordering and determine a best action, or (ii) accept that the choice is to some extent arbitrary, but look for a satisfactory, robust, reliable or convenient choice from amongst the maximal actions. We briefly outline some specific strategies, starting with type (i).

(a) Further assessment

Make more careful assessments, or use different assessment strategies, to try to make Your probabilities and utilities for Ω more precise.

(b) New possibility space

Try analysing the problem using a different possibility space Ω . Recall that Ω was introduced into the decision problem to help determine preferences. A new analysis, using a different Ω , may determine further preferences amongst the maximal actions.⁹

(c) Further information

Search for more information concerning Ω , to make Your probabilities more precise.

(d) Suspend judgement

Postpone a decision until a later time, when more information may be available. Where ‘suspending judgement’ is feasible it should be regarded as another option a_0 , to be compared with the other feasible actions on the basis of its associated costs and probabilities.¹⁰ Suspending judgement may often be a sensible response to indecision, but it should certainly not be identified with indecision.¹¹ In some problems where You are indecisive, it is simply not feasible to suspend judgement – an action must be taken immediately.

The above strategies are worth considering, but they will not always be successful in determining a unique action, especially when there is little available information and a decision cannot be postponed. Then one of the following methods may be used to select an action, without making any claim that this action is ‘best’.

(e) Satisficing¹²

Choose the first maximal action You find that is satisfactory. Often, some action will be distinguished by convenience, habit, convention, its conformity

with the status quo, or the predictability of its consequences. This action may be judged satisfactory if You have no clear preference for any alternative. The judgements of satisfaction can be formalized by defining the **satisfaction level** of an action to be the lower expected utility associated with it. When the utility function associated with an action is precise, it can be represented by a gamble X , and then the satisfaction level is simply $\underline{P}(X)$. This can be interpreted as the supremum price that You are disposed to pay in order to receive the consequences of the action. A satisfactory action can then be defined as one which achieves a pre-assigned satisfaction level.

(f) Minimax rules

Choose an action that maximizes the satisfaction level. In the case of precise utilities, this corresponds to choosing a gamble X that maximizes $\underline{P}(X)$, which is just the **\underline{P} -minimax rule** discussed in section 3.9.7.¹³

There are several ways of applying the \underline{P} -minimax rule to statistical problems. One way is to compare decision rules before obtaining the statistical data, by taking \underline{P} to represent prior beliefs about the parameter and $-X$ to be the risk function of a specific decision rule (as in section 3.9.8). A decision rule is chosen to have minimum prior upper expected risk.¹⁴

A second approach is to compare actions after obtaining the statistical data, by taking \underline{P} to represent posterior beliefs about the parameter and X to be the utility function associated with a specific action. An action is chosen to have maximum posterior over expected utility (or minimum posterior upper expected loss).

The decision rules generated by the second approach, by applying it to all possible statistical observations, are often non-maximal and therefore unreasonable. The first approach is more reliable; it always produces a maximal decision rule. More generally, when the \underline{P} -minimax rule is applied to several decision problems involving the same space Ω , it can generate decisions that are jointly unreasonable.¹⁵

(g) Precise probabilities and utilities

Choose a precise probability model and precise utility function that are consistent with the imprecise probabilities and utilities You have assessed, and then choose an action that maximizes expected utility.¹⁶ For purposes of decision making, this strategy is identical to the personalist Bayesian one. However, the precise probabilities and utilities are not models for Your ‘true’ beliefs and values, but merely a convenient way of reaching a decision in a specific problem. The imprecise models would be retained for purposes of inference and updating, and for decisions in other contexts.

5.6.7 Conclusion

The last three strategies are potentially useful, but they should not be regarded as procedures for establishing preferences between actions or for selecting an optimal action. They may be useful for selecting actions that are reasonable, satisfactory, robust or have other desirable properties, but we should not pretend that the action they lead to is the only reasonable action. Their usefulness seems likely to depend very much on the type and context of the decision problem, and on considerations of convenience and tractability, rather than on fundamental issues of rationality.

In the last resort, after carefully analysing the problem and using strategies (a) to (d) to eliminate all the unreasonable options, You can simply choose freely from amongst the remaining (maximal) options. This may be preferable to the arbitrary, and more difficult, choice of precise probabilities and utilities (g). (The choice of any other rule for resolving the indecision is also somewhat arbitrary.) Perhaps we should simply accept this degree of arbitrariness in choice, and call it ‘freedom of choice’, rather than follow rules to eliminate it. Reasoning cannot always determine a uniquely reasonable course of action, especially when there is little information to reason with.

5.7 Axioms of precision

Many axiom systems have been proposed as foundations for the Bayesian theory.¹ Such systems contain axioms of precision, which formalize the Bayesian dogma of precision. In this section we survey the various axioms of precision and the arguments put forward to support them.

The axioms of precision are all, essentially, axioms of completeness or comparability. They require You to rank the objects in a given class. The axioms differ according to the nature of the compared objects, which may be events, gambles, acts or randomized acts. The following four approaches are examined here.

1. **Fair prices** (Ramsey, de Finetti, Raiffa). It is assumed that Your supremum buying price for any event (or any gamble) coincides with Your infimum selling price. Essentially, this requires comparability of every event or gamble with every constant gamble.
2. **Comparative probability** (Jeffreys, de Finetti, Savage, Lindley, DeGroot). It is assumed that You can construct a complete comparative probability ordering of all events in some large class; for each pair of events, You are required to judge which is more probable.
3. **Preferences between acts** (Savage, Anscombe and Aumann, Fishburn). It is assumed that You can construct a complete preference ordering of all acts (or randomized acts) in some large class. (The class typically needs

to be much larger than the class of actions which are feasible in a decision problem.)

4. **Scoring rules** (de Finetti, Savage, Lindley). It is assumed that You can select the best gamble from amongst a large class of specified gambles (called ‘scores’ or ‘penalties’).

In considering these approaches, it is important to bear in mind the distinction between choice and preference (section 5.6.2). The Bayesian approaches require You to make various choices. If You make sufficiently many choices in a consistent way then a precise probability model can be constructed from them. But the resulting probabilities will reflect underlying beliefs only to the extent that the choices reflect underlying preferences. Probabilities derived from arbitrary or unstable choices are of little interest. It is therefore insufficient for Bayesians to argue that You can be induced to make choices and comparisons; they must explain why Your choices can be assumed to reflect genuine preferences.

5.7.1 Fair prices: de Finetti (1974)

De Finetti’s approach is especially interesting because, like the approach adopted in this book, it is based on a notion of coherence. The difference is that de Finetti assumes that upper and lower previsions (selling and buying prices) coincide for every gamble X . The common value $P(X)$ is called Your **fair price**, or simply **Your price**, for X . By Theorem 2.8.2, this assumption, together with the requirement that P avoids sure loss, forces P to be a linear prevision.

The key assumption is introduced by de Finetti in (1974, section 3.1.4): ‘We might ask an individual, e.g. You, to specify the *certain gain* which is considered *equivalent* to X . This we might call the *price* (for You) of X (we denote it by $P(X)$) in the sense that, on Your scale of preference, the random gain X is, or is not, preferred to a certain gain x according as x is less than or greater than $P(X)$.’ Thus he assumes that You can compare every gamble (random gain) X with every constant gamble (certain gain) x .

Why should Your preferences between random gains and certain gains be complete? De Finetti continues: ‘For every individual, in any given situation, the possibility of inserting the degree of preferability of a random gain into the scale of the certain gains is obviously a prerequisite condition of all decision-making criteria. Among the decisions which lead to different random gains, the choice must be the one that leads to the random gain with the highest price. Moreover, this is not a question of a condition but simply of a definition, since the price is defined only in terms of the very preference that it means to measure, and which must manifest itself in one way or another.’

It seems, however, that de Finetti is imposing a condition of completeness or comparability, and not simply a definition. In practice, You may have no preference between some gamble X and some constant x , without considering them to be equivalent. De Finetti is ruling out this possibility. But it is surely false that people have all the preferences he requires.²

Perhaps de Finetti does not mean that You already have these preferences, but only that You should try to construct them (although he gives little guidance on how to do that), because complete preferences are needed for making decisions. His argument seems to rest on the need for decisiveness, which was examined in section 5.6. In order to make a decision in a ‘given situation’, however, You do not need to compare all ‘the decisions which lead to different random gains’ with all the certain gains. These comparisons are needed in order to measure the prices $P(X)$, but much less is needed in order to compare the attainable random gains X with each other, still less is needed to choose a single random gain X_0 , and even the last choice does not imply a preference for X_0 over all the alternatives.

Thus de Finetti asks You to choose one of the attainable random gains X by choosing, for every such X , a price $P(X)$. Later he gives an operational definition of $P(X)$ as the value \bar{x} chosen by You when ‘You must accept a bet proportional to $X - \bar{x}$, in whatever sense chosen by Your opponent (i.e. positively proportional either to $X - \bar{x}$ or to $\bar{x} - X$).³ Here the precision of the elicited previsions is imposed by the elicitation procedure, and reveals nothing about whether or not Your underlying beliefs are indeterminate. Only rarely will there be a value \bar{x} which is both a buying and a selling price for X , such that You are willing (without compulsion) to accept either of the gambles $X - \bar{x}$ or $\bar{x} - X$. Again, choice need not reflect preference.

De Finetti does acknowledge that ‘all probabilities, like all quantities, are in practice imprecise’, but he emphasizes that the imprecision is not ‘essential’ and can always be reduced or eliminated by further thinking and calculation.⁴

5.7.2 Fair betting rates: Ramsey (1926)⁵

Ramsey’s approach is similar to de Finetti’s, although Ramsey gives more attention to utility and rather less attention to probability than de Finetti does. After constructing a linear utility scale for gambles, Ramsey in effect defines Your ‘degree of belief’ in an event A as the real number α for which You are ‘indifferent’ between the gamble A and the constant gamble α (1926, p. 179). ‘Indifferent’ has the same meaning as ‘equivalent’ does for de Finetti, and it is implicitly assumed that preferences between events and constant gambles are complete. Hence α is Your ‘fair price’ for the gamble A , in de Finetti’s sense.⁶

Ramsey states his fundamental assumption with admirable clarity:

'I propose to take as a basis a general psychological theory, which is now universally discarded, but nevertheless comes, I think, fairly close to the truth in the sort of cases with which we are most concerned. I mean the theory that we act in the way we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions.' (1926, p. 173). This rules out indeterminacy in values and beliefs, which could lead to indecision. Since people do have indeterminate beliefs and do make arbitrary choices, Ramsey's 'psychological theory', which seems to be intended as a description of human behaviour, seems inadequate.

In fact Ramsey does recognize two sources of imprecision in probabilities: 'First, some beliefs can be measured more accurately than others; and, secondly, the measurement of beliefs is almost certainly an ambiguous process leading to a variable answer depending on how exactly the measurement is conducted.' (1926, p. 167). But he regards probability as no different in these respects from other physical concepts (such as time intervals), and he makes no allowance for imprecision in his theory.

5.7.3 Comparative probability: Jeffreys (1931, 1983)

Many formal theories of probability have been based on axioms for a comparative probability ordering of events or propositions.⁷ In these theories, a completeness axiom is needed to obtain precise numerical probabilities whose order agrees with the comparative probability ordering. Jeffreys' first axiom is a completeness axiom, in which p , q , r are propositions: 'Given p , q is either more, equally, or less probable than r , and no two of these alternatives can be true.' (1983, p. 16).

After adding further axioms, such as transitivity of the ordering, Jeffreys assigns real-valued probabilities to propositions conditional on data. These agree with the ordering in the sense that larger numbers are assigned to more probable propositions, but Jeffreys regards the assignment of numbers as merely conventional; only the ordering is significant.

The only justification given by Jeffreys for his completeness axiom is that 'It is generally believed that probabilities are orderable... Even if people disagree about which is the more probable alternative, they agree that the comparison has a meaning.' (1983, pp. 15–16). Jeffreys does not explain the meaning of the comparison, however. His reticence is surprising because earlier authors, including Keynes (1921), who adopted a similar logical interpretation of probability and whose work was studied by Jeffreys, had denied that probabilities could be completely ordered.⁸

Jeffreys also gives little guidance on how to recognize that one proposition is more probable than another on given data p , except in the case where

p is a tautology. In that case he advocates a version of the principle of insufficient reason, which leads to well-known inconsistencies (section 5.5.1).⁹ The completeness axiom seems unreasonable in the case where p provides little or no information with which to compare q and r .

5.7.4 Comparative probability: Lindley (1971a)

Like Jeffreys, Lindley bases his account of probability on judgements of comparative probability, although Lindley adopts a personalist rather than a logical interpretation. Lindley's fundamental assumption is essentially the same as Jeffreys' completeness axiom, except that Lindley refers to events rather than propositions: 'The argument really rests on the comparability of any two uncertain events... This assumption of comparability must therefore be considered carefully.' (1971a, p. 21).

Lindley gives three arguments to support this assumption. His first argument is that incomparable events must be 'of different types': 'If not all uncertain events can be compared there must be at least two types of uncertain event: events of any one type may be related, but not events of different types.'¹⁰ This seems mistaken. Events of the same 'type' may be incomparable, for instance the events that two different thumbtacks land pin-up when tossed, if there is little information on which to base a comparison. Contrary to Lindley's suggestion, a partial ordering need not be the superposition of complete orderings each involving 'one type' of events.

Lindley's second justification is 'the observable fact that some events which seem unpromising candidates for numerical assessment of probabilities are so assessed, even if the assessments are only made indirectly' (1971a, p. 23). He discusses two examples: horse-racing and insurance. In the case of horse-racing, 'It seems hard to estimate a probability and yet bookmakers will quote odds...'. This is a curious example since it is well known that bookmakers' odds correspond to non-additive upper probabilities. Bookmakers aim to 'make a book', thereby guaranteeing themselves a sure gain, and they cannot do so by quoting additive probabilities.¹¹

In Lindley's second example, You must decide whether or not to take out insurance against an event at a specified premium. 'A comparison must be made: there is no way out.' This is the argument for decisiveness (section 5.6). You may indeed need to choose, but the choice may not be determined by Your beliefs about the event, especially when You have no statistical information about the frequency of similar events. For some range of insurance premiums, You may be unable to judge whether insurance is worthwhile.¹²

Lindley's third argument (1971a, pp. 113–16) is that the 'vagueness' we associate with some probability assessments can actually be modelled in terms of precise probabilities. He compares the event that a new (untested) drug will do better in a test than an old drug, with the event that an ordinary coin will land 'heads' on its next toss.¹³ Lindley suggests that You might assess precise probability $\frac{1}{2}$ for each event, but regard the probability for the coin as 'firmer' or 'less vague' than that for the drug. The difference in vagueness can be reflected in: (a) Your probabilities for the outcomes of a sequence of similar trials, or (b) Your updated probabilities after observing the outcomes of earlier trials. Of course the vagueness is not reflected in the probabilities of the two events themselves, which are both precisely $\frac{1}{2}$, and Lindley claims that it is 'irrelevant for the decision in hand'.

Consider assessing the probability of a unique event, such as the meltdown of a nuclear power plant (the only one of its type) during its lifetime. There might be considerable vagueness associated with Your assessment, and it is unnatural to explain this in terms of 'similar trials' which will never be observed. According to Lindley, the vagueness is irrelevant to a decision about whether to close down the plant.¹⁴ You are told to assess a precise probability of meltdown (0.00031562..., or whatever) and use this to reach a clear-cut decision. But any precise assessment would be arbitrary, and its precision would be spurious. 'Vagueness' or lack of information should be reflected in imprecise probabilities, perhaps leading to indecision about whether to close the plant. It is sometimes important to acknowledge our ignorance.

Finally, Lindley acknowledges that we cannot always make probability comparisons in practice (1971a, pp. 170–1), but he regards this as a problem of 'application' or 'technology' rather than a defect of the theory. But what 'technology' will enable us to assess reasonable, precise probabilities when we have little information? We have argued that any method for constructing precise probabilities from ignorance will lead to difficulties. Many of the difficulties in applying the Bayesian theory come from the fundamental requirement of precision, and these cannot be overcome by merely 'technological' developments.

5.7.5 Preferences between acts: Savage (1972a)

In Chapter 3 of his book, Savage derives precise probabilities from de Finetti's axioms for comparative probability (Appendix B5) together with a partitioning axiom. The objectionable axiom is Savage's axiom P4, which requires completeness of the comparative probability ordering.¹⁵ In order to derive also a precise utility function, Savage needs a stronger axiom (his P1), which requires completeness and transitivity of a preference relation between acts.¹⁶ Acts are functions mapping the possibility space Ω into a

space of consequences. They may be interpreted as imaginary actions resulting, under each possible state of affairs, in a specified consequence. Not all such actions will be feasible in a given decision problem.

Savage does not argue directly for his axiom P4, but he does discuss his stronger axiom P1 (1972a, pp. 19–21). He claims that P1 is 'normative' and 'can be regarded as a logic-like criterion of consistency in decision situations'. Axiom P1 requires the ordering of acts to be both complete and transitive. Savage gives a detailed argument to support transitivity. Intransititivity of preferences is irrational because it leads, through the 'money-pump' argument (section 1.6.3), to a sure loss. But what is irrational about incompleteness? In fact, Savage gives no argument to support completeness, and his brief comments suggest that he did not regard completeness as normative in the way that transitivity is.

Savage was certainly aware that the precision demanded by the Bayesian theory is often unrealistic, and he refers to this in many of his writings. See, in particular, 'Some shortcomings of the personalistic view' in Savage (1972a, Ch. 4).¹⁷

5.7.6 Scoring rules: de Finetti (1974)

De Finetti has proposed a second 'operational' definition of previsions, based on the quadratic scoring rule.¹⁸ Suppose that You are concerned with the gambles X_1, X_2, \dots, X_n . To elicit Your previsions for these gambles, we require You to specify real numbers x_1, x_2, \dots, x_n . After the true state ω is observed, You are required to pay a penalty $\sum_{j=1}^n (X_j(\omega) - x_j)^2$ in units of utility. In other words, You must choose to accept one of the gambles $S(\underline{x}) = -\sum_{j=1}^n (X_j - x_j)^2$ determined by Your choice of $\underline{x} = (x_1, x_2, \dots, x_n)$.

The chosen \underline{x} is 'admissible' or 'coherent', according to the criterion of de Finetti (1974, section 3.3.6), when there is no alternative choice which would produce a uniformly lower penalty. De Finetti shows that a response \underline{x} is admissible if and only if it defines a linear prevision P by $P(X_j) = x_j$. For example, if $X_3 = X_1 + X_2$ but $x_3 \neq x_1 + x_2$ so that the linearity property is violated, You can obtain a uniformly smaller penalty by modifying the choices x_1, x_2, x_3 to $x'_1 = x_1 + \delta$, $x'_2 = x_2 + \delta$ and $x'_3 = x_3 - \delta$, where $\delta = (x_3 - x_1 - x_2)/3$.¹⁹ This shows that a prevision P elicited through the quadratic scoring rule should be linear.

Our reply to this argument is that it again confuses choice with preference. In the particular decision problem involving the quadratic scoring rule, Your choice \underline{x} should correspond to some linear prevision P because any other choices will be inadmissible. But there may be some arbitrariness in the choice of a linear prevision P , as this need not be determined by beliefs. If Your beliefs are indeterminate, modelled by a lower prevision \underline{P} , then it

is maximal under P to choose any linear prevision P in $\mathcal{M}(\underline{P})$. The argument does not establish that Your beliefs are, or should be, determinate, since the linearity of the response \underline{x} is determined not by Your beliefs, but by the mathematical form of the scoring rule.²⁰

To illustrate this, consider a different type of scoring rule, the **absolute-error** rule. For each subset A_j of a finite space Ω , You must specify Your ‘probability’ x_j . You then receive the gamble $S(\underline{x}) = -\sum_{j=1}^n |A_j - x_j|$. Suppose, for simplicity, that Your beliefs really can be represented by a linear prevision P . Then $P(|A_j - x_j|) = |1 - x_j|P(A_j) + |x_j|P(A_j^c)$, and this is minimized by choosing $x_j = 1$ if $P(A_j) > \frac{1}{2}$, $x_j = 0$ if $P(A_j) < \frac{1}{2}$, and any value x_j between 0 and 1 if $P(A_j) = \frac{1}{2}$.

It follows that, under P , the maximal choices \underline{x} are the **classificatory** responses, where You choose $x_j = 1$ if You judge A_j to be probable, and $x_j = 0$ if A_j is improbable.²¹ Of course, this response does not mean that You are only capable of classifying events as probable or improbable. Just as with the quadratic scoring rule, the type of response is determined by the mathematical form of the rule rather than by the precision of Your underlying beliefs.

The absolute-error scoring rule can be useful when Your beliefs really are of the crude classificatory type. Similarly, the quadratic scoring rule is useful when Your beliefs really are determinate. But the existence of these scoring rules tells us nothing about what structure Your beliefs do, or should, have.²²

5.7.7 Conclusion

We have examined various arguments put forward by Bayesians to support their axioms of precision. None of these arguments is at all convincing. (Compare with the rationality arguments for avoiding sure loss and coherence in Chapters 2, 6 and 7, which we do find convincing.) The objectionable axioms all require complete comparability of events, gambles or acts.²³ The arguments advanced to support comparability show only that You can be forced to make choices, not that You should have beliefs, values and preferences which fully determine these choices. None of the theories shows You how to make the many comparisons needed to determine precise probabilities. Since the various arguments for precision have essentially the same weakness, they remain unconvincing when viewed together. We conclude that the axioms of precision are unjustified.

5.8 Practical reasons for precision

Next we examine some of the pragmatic arguments for precise probabilities. It is common for Bayesians to admit that there are practical difficulties

in achieving precision, but to argue that precise probabilities are an ‘idealization’ or ‘ideal’ at which You should aim. Other arguments are that imprecise probabilities are too difficult to assess or too complicated to work with, and that Bayesian methods of statistical inference are perfectly adequate in practice.

5.8.1 Comparison with other scientific concepts

It is sometimes claimed that precise probabilities are an ‘idealization’, in the same way that Euclidean geometry or Newtonian mechanics are idealizations of physical reality.¹ However, different sorts of idealization seem to be involved in these three cases. Euclidean geometry is an idealization in that there are practical limits to the precision with which real distances and angles can be measured. Similarly there are limits to the precision with which real beliefs can be measured. The analogy quickly breaks down, however, because the imprecision in probabilities arises more from indeterminacy in beliefs than from measurement errors.

The concept of probability in the Bayesian theory may be compared more aptly with the concept of the time interval between two instantaneous events in the Newtonian theory. It is known, from the theory of relativity, that a time interval is not precisely determined. The Newtonian theory, in which time intervals are precise, is not an exact description of reality. Similarly the Bayesian theory, in which probabilities are precise, is not an exact description of real beliefs, which are typically indeterminate. Nevertheless, Ramsey (1926) argues that, for both time intervals and probabilities, the idealization of precision is ‘sufficiently accurate for many purposes’.

There are several responses to these comparisons between probability and other scientific concepts.

1. It is unclear whether Bayesians intend the axiom of precision to be an ‘idealization’ or an ‘ideal’. Comparisons with Euclidean geometry and Newtonian mechanics, which are both used to describe real phenomena, suggest that the Bayesian theory of probability is viewed as a description of human behaviour. For the comparisons to be appropriate, the indeterminacy in real beliefs must be negligible for practical purposes. That is evidently not so. Indeed, most Bayesians seem to regard precision as normative (an ideal) rather than descriptive (an idealization). They argue that, in order to be rational, You *should* obey the axioms of precision. It would be absurd to argue that real phenomena should obey the Euclidean axioms or the Newtonian laws.²
2. In applications of Euclidean geometry or Newtonian mechanics where there is appreciable measurement imprecision, some kind of error theory

or sensitivity analysis is commonly used to assess the effect on conclusions.³ It is crucial in structural engineering, for example, to model the propagation of measurement errors and to allow sufficient tolerances for these. The standard measurement procedures for physical quantities such as length, time and mass include procedures for obtaining bounds on measurement errors. These directly indicate to what extent the idealization of precise measurement is a reasonable one. For probability, on the other hand, it is not standard practice to assess bounds for measurement errors or imprecision, which would be tantamount to assessing upper and lower probabilities. Without these, it is difficult to judge whether the idealization of precision is reasonable in a specific application.

3. Measurement errors are not the only source, or even the major source, of imprecision in probabilities.⁴ The high degree of imprecision of the near-ignorance models in section 5.3 is due to lack of information rather than measurement error. Similarly, the incompleteness in our ordering of instantaneous events may be due to relativistic effects rather than measurement error. Just as we need the theory of relativity to tell us when these effects are important and the Newtonian theory is a poor idealization, so we need a theory of imprecise probabilities to tell us when precision is a poor idealization.

4. An important practical difference between probabilities and time intervals is that the indeterminacy in time intervals is negligible for ‘ordinary’ events (where relativistic effects are negligible), whereas the degree of imprecision in probabilities is ordinarily appreciable. When probabilities are based on little information, even the first decimal place may be undetermined. In such cases the ‘idealization’ of precision is seriously misleading.⁵ The analogy with the ordering of events is closer when we consider events with duration, such as a person’s lifetime, as well as instantaneous events, such as their death. Some events can be located at a precise point in time, but for other events this would be an absurd idealization. Probabilities, like occurrence times, often need to be represented by intervals rather than points.

5.8.2 Two precise numbers instead of one

A common objection to upper and lower probabilities is that their specification requires two precise numerical assessments, $\bar{P}(A)$ and $\underline{P}(A)$, for every one, $P(A)$, that is needed by Bayesians.⁶ A theory of imprecise probabilities seems to demand more precision, and greater effort in assessment, than the Bayesian theory! It should be clear from the following considerations that this objection is mistaken, at least under the interpretation of upper and lower probabilities that is adopted here.

1. The assessments $\bar{P}(A)$, $\underline{P}(A)$ and $P(A)$ are not merely numbers; they are models for Your betting rates, and other behavioural dispositions, concerning the event A . When $\underline{P}(A) < P(A) < \bar{P}(A)$, the Bayesian model which asserts that You have two-sided betting rate $P(A)$ is making stronger claims about Your behaviour than the imprecise probability model. It should be easier for You to assess acceptable upper and lower betting rates $\bar{P}(A)$ and $\underline{P}(A)$ than to assess an acceptable two-sided rate $P(A)$.
2. In the general elicitation procedure of section 4.1, any judgements You make will generate upper and lower probabilities, but it is much more difficult to obtain precise probabilities (section 4.1.6). For example, You can construct upper and lower probabilities by assessing a comparative probability ordering of subsets of Ω . The resulting model will be precise only in unusual cases, and difficult comparisons with extraneous measurement events will usually be needed to achieve precision (section 4.5.7).
3. When there is an underlying precise probability $P(A)$ which You are trying to assess or elicit, it is clearly more difficult to assess $P(A) = 0.37954\dots$ precisely than to say that $0.3 \leq P(A) \leq 0.5$. The difference is analogous to the difference between measuring the length of an object as precisely 0.37954... centimetres, and bounding the length between 0.3 and 0.5 centimetres. The imprecise measurement does involve two precise numbers instead of one, but it is certainly easier to make than a precise measurement.
4. If Your behavioural dispositions really are indeterminate, and exhaustively modelled by upper and lower probabilities $\bar{P}(A)$ and $\underline{P}(A)$, then it may be about as difficult for You to elicit each of the numbers $\bar{P}(A)$ and $\underline{P}(A)$ as for a Bayesian to elicit his precise probability $P(A)$. The objection of ‘two precise numbers instead of one’ therefore has some force against an exhaustive interpretation (section 2.10.3), although not against our interpretation which requires You to assess only upper and lower bounds for Your exhaustive $\bar{P}(A)$ and $\underline{P}(A)$.⁷
5. The class of coherent lower previsions defined on some space Ω is very much larger than its subclass of linear previsions, and in that sense a lower prevision is more difficult to select than a linear prevision.⁸ Suppose that Ω has finite cardinality n . Then a linear prevision is completely determined by the probabilities of $n - 1$ elements of Ω , so the class of linear previsions has dimension $n - 1$. The class of coherent lower probability models, defined on all subsets of Ω , has dimension $2^n - 2$.⁹ The class of coherent lower previsions on $\mathcal{L}(\Omega)$ is even larger, since lower previsions are not determined by lower probabilities, and actually has infinite dimension when $n \geq 3$.

Thus the number of coherent models is very greatly increased when imprecision is admitted. You can, of course, restrict the choice of an imprecise model P to a much smaller class. Bayesians often choose their models from small classes of tractable models, such as conjugate priors, each specified

by a few real parameters. Similarly, P can be chosen from models of a particular form, such as the classes of conjugate priors defined in section 5.3.

5.8.3 Imprecise probabilities are complicated or intractable

This objection is partially answered in the previous paragraph: most imprecise probability models are intractable in real problems, just as most Bayesian models are, but tractable models of both kinds can be developed. Much work needs to be done to develop simple, tractable types of imprecise models, but the enormous gain in realism from admitting imprecision is enough to justify this effort.¹⁰

It is true that the theory of additive probabilities and linear previsions is mathematically simpler than the general theory developed in this book. (Linearity is simpler than super-linearity.) We would argue, however, that imprecise probabilities are conceptually simpler. Upper and lower previsions can be easily understood as selling and buying prices for gambles, and it is natural to suppose that these prices differ. They are certainly easier to assess or elicit than precise probabilities, through judgements of classificatory or comparative probability (for instance) that can be easily made without training in probability theory.

In any case, since much of the theory of upper and lower previsions can be interpreted as a formalization of sensitivity analysis, the mathematical complications introduced by admitting imprecision need be no greater than those accepted by Bayesian sensitivity analysts. The introduction of upper and lower previsions can actually simplify sensitivity analysis by clarifying its behavioural meaning and providing an alternative mathematical calculus.

5.8.4 Arguments for Bayesian inference

The Bayesian theory of statistical inference is often defended by pointing to its advantages over other statistical methods.¹¹ The most convincing advantages are the following:

1. The Bayesian approach ensures consistency (coherence) amongst potential inferences and decisions.
2. In statistical problems we need to measure uncertainty about the quantities of interest after observing statistical data. That is, we need posterior measures of uncertainty, which are conditional on the data.
3. It is often useful to formally model prior information concerning statistical parameters.
4. In statistical decision problems it is reasonable to restrict attention to the admissible decision rules, which are essentially the Bayes' rules

obtained from different prior distributions. Preferences amongst the admissible rules are therefore equivalent (in effect) to judgements of prior probabilities.¹²

5. The Bayesian approach is consistent with the likelihood principle, whereas frequentist statistical methods are not.
6. 'Statistical inference' is a special type of reasoning under uncertainty, which should be covered by a general theory.

These six arguments are persuasive arguments against frequentist methods, but they should not persuade us to become Bayesians. The arguments support the approach to statistical inference that is developed in Chapters 7 and 8, which is a special case of our general theory, but they do not support the Bayesian dogma of precision. Prior information can be modelled by imprecise prior probabilities, leading to imprecise posterior probabilities and coherent inferences. You may only be able to partially order the admissible decision rules. (Even a complete ordering may not determine precise prior probabilities.) The likelihood principle is compatible with imprecise prior and posterior probabilities; indeed, coherent inferences based on imprecise priors will automatically satisfy the likelihood principle in the discrete case (section 8.6).

5.8.5 The Bayesian theory 'works'

A final argument for the Bayesian approach is that 'it works', meaning that it can be used to reach sensible conclusions in practical problems. If that were so then there would be no need, in practice, to use imprecise probabilities. Our view, of course, is that there are many practical problems where the Bayesian approach does not work because precise probabilities cannot be justified.

It is not entirely clear, however, what is intended by the claim that the Bayesian approach leads to sensible conclusions. Conclusions are not determined simply by applying the Bayesian theory – they will vary according to what prior probabilities are assessed. If the prior assessments are arbitrary, as allowed by personalist Bayesians, then the conclusions will be arbitrary. So the Bayesian approach can 'work' only if it includes methods for determining sensible prior probabilities. Bayesians seem unable to find such methods, at least in problems where there is little prior information (section 5.5).

5.9 Bayesian sensitivity analysis

The conclusion we draw from the preceding part of this chapter is that it is often reasonable to hold indeterminate beliefs and to model these through

imprecise probabilities. In the rest of the chapter we examine some alternative ways of doing this.¹ We start with Bayesian sensitivity analysis, which seems to be the most popular way of modelling indeterminacy and is also the closest to our approach.²

5.9.1 *The dogma of ideal precision*

Bayesian sensitivity analysis replaces the single probability measure P used in standard Bayesian analysis by a class \mathcal{M} of probability measures. It is assumed that there is some true or ideal probability measure P_T , but You are uncertain about what this is. So \mathcal{M} is a set of possible (or plausible) candidates for the true probability measure.

One motivation for sensitivity analysis is the lower envelope theorem 3.3.3, which asserts that a lower prevision \underline{P} is coherent if and only if it is the lower envelope of some class of linear previsions. This result is consistent with a sensitivity analysis interpretation, since whenever \underline{P} is coherent, it could arise from a sensitivity analysis in which the analyst could determine only that the true probability P_T belonged to the set $\mathcal{M}(\underline{P})$.

There are two basic questions concerning sensitivity analysis: what is the meaning of the ‘true’ or ‘ideal’ P_T ; and is it reasonable to assume that such ‘true precise probabilities’ exist? It is essential to answer these questions, because when You construct a class \mathcal{M} You are, according to sensitivity analysts, expressing beliefs concerning the true P_T . To do so, You need to define the precise probability P_T , and You need to be sure that there is a unique P_T which satisfies this definition, even though You are uncertain about what it is.

Some possible interpretations for P_T were outlined in section 2.10.4. These may be classified as descriptive or normative. According to the descriptive interpretations, Your beliefs actually are determinate and could be properly modelled by a precise P_T , except that P_T cannot be precisely elicited because of the difficulties discussed in section 5.2. This answers the first question clearly enough: the true P_T is the correct model for Your real (determinate) beliefs. But then the answer to the second question must be negative. There is plenty of psychological evidence that personal beliefs are not both determinate and coherent.³

The normative interpretations seem to be based on the **dogma of ideal precision**: precise probability assessments are an ideal which should be aimed at but which cannot be attained in practice, due to limitations of time or analytical ability, or other practical difficulties. The ideal probability measure P_T is the one that would result from an ideal process of assessment, if You had the time and ability to carry this out. The normative interpretation is much vaguer than the descriptive interpretation, because it is not at all

clear what an ‘ideal process of assessment’ is supposed to involve. Nor is it clear why it should result in precise probabilities.

The two basic questions are not well answered in the literature on Bayesian sensitivity analysis. Let us examine the answers given by Good and Berger, the two sensitivity analysts who have considered most carefully the foundations of their theory.

5.9.2 *Black boxes: Good (1962a)⁴*

Good’s ‘black box’ is a formal model for Bayesian sensitivity analysis. Good recognizes that probability judgements are usually imprecise. All Your imprecise judgements are fed into a formal system called a black box, which translates them into constraints on a precise probability P_T . For example, the judgement that event A is more probable than B is translated into the constraint $P_T(A) > P_T(B)$.⁵ These constraints are then combined with the Bayesian axioms for precise probabilities to produce new constraints, such as $P_T(C) > P_T(D)$, which can be translated back into discernments, such as ‘ C is more probable than D ’. The output of the black box consists of such discernments, which are added to Your body of beliefs. In effect, the black box uses Your judgements to determine a class \mathcal{M} of precise probabilities (all those satisfying all the constraints), each of which it manipulates in the standard Bayesian way.

Good’s key assumption is that there is an unobservable, precise P_T inside the black box.⁶ For instance, upper and lower probabilities are regarded as upper and lower bounds for the precise P_T . But how is this assumption to be justified? What is the meaning of the unobservable P_T , and why should we assume it exists?

Good does not directly answer these questions, and perhaps the image of a ‘black box’ is a deliberate attempt to remain non-committal on the meaning of P_T . In other writings, Good advocates a subjective (personalist) interpretation of probability, but he suggests that it is often useful to regard personal probabilities as estimates of logical probabilities.⁷ On that view, we might interpret the P_T inside the black box as the correct (but unknown) logical probability measure, and \mathcal{M} as the set of measures You judge to be plausible candidates. Alternatively, we might regard Your mind as a black box. Then P_T describes Your real but unobservable beliefs. Neither of these interpretations justifies the assumption that probabilities inside the black box are precise.

Elsewhere, Good claims that ‘A justification of this extremely simple black-box approach... has been given by Smith (1961, 1965).’⁸ Apparently this refers to the lower envelope theorems such as Theorem 3.3.3. But these results show only that a coherent, unconditional, imprecise probability

model is mathematically equivalent to a class of precise probability measures. These theorems do not confer any meaning on the precise probability measures, so they do not support the assumption that one precise measure P_T is the ‘true’ measure.

Finally, Good acknowledges that his assumption ‘is like saying of a non-measurable set that it really has an unknowable (“metamathematical”) measure lying somewhere between its inner and outer measures’ (1962a, p. 324). Of course, this assumption is not needed and not accepted by measure theorists. The corresponding assumption concerning upper and lower probabilities is also unnecessary.

5.9.3 Robust Bayes: Berger (1984)

An especially lucid account of the foundations of Bayesian sensitivity analysis is presented by Berger (1984). His approach is to assess a class \mathcal{M} of ‘plausible’ precise probabilities, and similarly a class of precise utility functions, and carry out a standard Bayesian analysis with each pair of precise probabilities and utilities. The analysis is ‘robust’ if the conclusion or decision is approximately the same for each pair. According to Berger (1984), the robust Bayesian viewpoint ‘is essentially that one should strive for Bayesian behaviour which is satisfactory for all prior distributions which remain plausible after the prior elicitation process has been terminated’.

What does Berger mean by ‘plausible’ prior distributions? Apparently he means plausible candidates for the ideal prior distribution P_T , which is ‘the prior which would be the result of infinitely long reflection on the problem’.⁹ This imaginary prior ‘is exactly nailed down only after an infinite process of elicitation’.¹⁰ Thus Berger adopts a personalist version of the dogma of ideal precision: You just do not have the time to assess the ideal probabilities! Indeed, the fundamental Assumption 2 of Berger (1984) states: ‘Prior distributions can never be quantified or elicited exactly (i.e. without error), especially in a finite amount of time.’

Berger answers the first question of section 5.9.1 by interpreting P_T as the probability measure You would adopt if You had infinite time for assessment. But is this interpretation sufficiently definite for You to judge, in practical problems, which measures are ‘plausible’, i.e. which measures You plausibly might adopt if You had infinitely long to think about it? His characterization of P_T would be more useful if Berger held a logical interpretation of probability and could provide a method which uniquely defined the ‘ideal’ P_T , but there is no such method. It is unclear what kinds of judgement and analysis should be included in this ‘infinitely long reflection’.

In any case, Berger does not answer the second question of section 5.9.1:

5.9 BAYESIAN SENSITIVITY ANALYSIS

why should You settle on any precise probability model, even after an infinite amount of reflection? Evidently he regards lack of time in elicitation or assessment as the main source of imprecision in probabilities. But there are many other sources of imprecision, some of which are more important (see section 5.2).¹¹

Consider the case of little or no prior information, for example. Why should You need much time to analyse no information? ‘Infinitely long reflection’ certainly seems excessive! Yet Bayesian noninformative priors are inadequate models in this case, and highly imprecise probabilities are needed. Berger himself gives an example to show that ‘there are situations in which it seems simply unreasonable to expect that beliefs can even be modeled [ideally] by a single prior distribution.’ (1984, p. 71).

5.9.4 Conclusion

Both Good and Berger assume that there is an unobservable, ideal probability model P_T which is precise. Neither justifies this assumption. We have argued that, far from being ‘ideal’, precision may misrepresent the available information. If no precise probability can be justified even through an ideal analysis of the evidence, there seems little sense in regarding the individual probability measures in \mathcal{M} as plausible candidates to be the ‘ideal’ measure. We therefore see no reason to adopt a sensitivity analysis interpretation of imprecise probabilities.

The attempts to establish a sensitivity analysis interpretation, besides being artificial and unconvincing, are simply unnecessary. We can develop an adequate theory of probability without them, based only on the behavioural interpretation and appropriate concepts of coherence.

Whereas sensitivity analysts emphasize the class \mathcal{M} of precise probabilities, we would emphasize its upper and lower envelopes \bar{P} and \underline{P} . The difference in emphasis has important practical consequences (section 2.10.5). In particular, we would assess a prior probability model $\mathcal{M}(\underline{P})$ by examining its behavioural implications, expressed through properties of \bar{P} and \underline{P} , rather than by trying to determine which precise models P are ‘plausible’, through properties of the individual P in $\mathcal{M}(P)$. When there is little prior information, for example, we would require \underline{P} to have properties such as near-ignorance (section 5.3).¹²

We must stress, however, that there are great mathematical similarities between our approach and sensitivity analysis, because of the lower envelope theorems such as Theorem 3.3.3. There are also practical similarities. Many of the specific models proposed by sensitivity analysts can be useful under a behavioural interpretation. The inference method used by sensitivity analysts, which involves selecting a class \mathcal{M} of prior distributions, updating

each of these by Bayes' rule, and drawing those conclusions that are valid under all of the resulting posterior distributions, is a coherent and sensible one. In most problems we would use the same inference strategy, but our conclusions will often differ from those of sensitivity analysts because we would use different criteria in selecting the prior class \mathcal{M} .¹³ Our approach is certainly closer to Bayesian sensitivity analysis than to the more elaborate approaches discussed in the following sections.¹⁴

5.10 Second-order probabilities

In Bayesian sensitivity analysis, the linear previsions in the class \mathcal{M} are regarded as possible models for ideal beliefs concerning Ω . That is, \mathcal{M} is regarded as a possibility space. You may have non-vacuous beliefs concerning the possibility space \mathcal{M} ('beliefs about beliefs'), which can be expressed in the form of **second-order probabilities** on \mathcal{M} . It is therefore natural to consider a theory of second-order probabilities as an elaboration of sensitivity analysis.¹

5.10.1 Basic model

Suppose that You are interested in the possibility space Ω , and You use the elicitation procedure of section 4.1 to construct a class \mathcal{M} of linear previsions, each defined on $\mathcal{L}(\Omega)$. Let P_2 be a linear revision defined on $\mathcal{L}(\mathcal{M})$, or more generally on $\mathcal{L}(\mathcal{P})$, representing Your current beliefs about the 'true' or 'ideal' probabilities P_T . (The subscript 2 indicates that these are second-order probabilities.)

A theory of second-order probabilities inherits the fundamental problem of sensitivity analysis: it needs to give a clear interpretation of the 'true' or 'ideal' P_T , before You can hope to assess probabilities concerning P_T . Whatever interpretation is adopted, it seems reasonable to regard the possibilities P in \mathcal{M} also as conditional previsions, in the sense that if You learned that $P_T = P$, then You would adopt the behavioural dispositions determined by P . We assume also that P_2 can be given a behavioural interpretation, in terms of Your betting rates concerning P_T , however this is defined.

5.10.2 First-order probabilities

Composite first-order probabilities P_1 , representing marginal beliefs about Ω , can be constructed by natural extension.² The linear revision P_1 is defined on $\mathcal{L}(\Omega)$ by $P_1(X) = P_2(X^*)$, where X^* is the gamble defined on \mathcal{M} by $X^*(P) = P(X)$. Thus $P_1(X)$ is the revision (under the second-order distribution P_2) of the unknown first-order prevision $P_T(X)$.

Provided P_2 and the previsions in \mathcal{M} have the behavioural interpretations mentioned above, they effectively imply, through coherence, the behavioural dispositions represented by P_1 . In order to satisfy coherence, You must adopt the precise probabilities P_1 as Your current betting rates concerning Ω . Thus the imprecise class \mathcal{M} is reduced, through the assessment of P_2 , to the single linear revision P_1 , and the imprecision is eliminated.³

This result may seem somewhat paradoxical. We started by assuming that You were uncertain about Your 'true' linear prevision P_T but concluded that You should adopt the linear prevision P_1 . To avoid incoherence, it is necessary to give different interpretations to P_T and P_1 . The 'true' prevision P_T cannot be interpreted as a model for current behavioural dispositions, since P_1 has this interpretation.⁴

5.10.3 Aleatory and logical interpretations

In some problems P_T can be given a frequentist or propensity interpretation. In statistical problems, for example, \mathcal{M} may be a set of sampling models, exactly one of which (P_T) generates the statistical data. Then P_2 is a Bayesian prior distribution, and P_1 is simply the predictive distribution for the statistical observation.⁵

In other problems, the unknown P_T might be interpreted as the correct logical prevision which would result from an ideal analysis of the available evidence concerning Ω . Then P_2 represents Your current beliefs about the unknown logical probabilities, and P_1 represents Your current behavioural dispositions concerning Ω .⁶

Suppose, for example, that ω is the tenth digit in the decimal expansion of π , so that $\Omega = \{0, 1, \dots, 9\}$. Since ω could be determined, in principle, from an elementary knowledge of mathematics, the possible logical probability distributions are the ten degenerate distributions on Ω . An ignorant Bayesian might take P_2 to be uniform on the space \mathcal{M} consisting of the ten degenerate distributions. The first-order probability distribution P_1 is then uniform on Ω . Until he makes the necessary calculations to determine ω , the Bayesian's would assign probability 0.1 to each of the ten possible unless.

In general it is not clear how to define the correct logical probabilities, and nor is it clear why they should be precise. There appear to be similar problems with other interpretations of P_T . It is somewhat difficult to assess second-order probabilities concerning quantities whose definition is unclear.

5.10.4 Bayesian hierarchical models⁷

Suppose that ω is a statistical parameter that indexes a class of hypothetical sampling models. Bayesians need to assess a precise prior distribution for

ω . That might be done in two stages, by first assessing a class \mathcal{M} of ‘possible’ prior distributions for ω , usually indexed by some hyperparameter ϕ , and then assessing a hyperprior P_2 for ϕ . As in section 5.10.2, this model generates a composite prior P_1 for ω .

Such hierarchical models can be very useful when the hyperparameter ϕ has a clear meaning, so that the ideal prior P_T indexed by ϕ can be given an aleatory or logical interpretation. But the hyperparameter seems to be introduced, in many Bayesian analyses, merely as an index for the unknown, uninterpreted P_T . It is difficult to see how a hyperprior can be realistically assessed in such cases. Indeed, Bayesians commonly model their ‘complete ignorance’ about ϕ through a ‘noninformative’ hyperprior P_2 . The dangers of such models are apparent from section 5.5.⁸

5.10.5 Imprecision

It seems necessary to allow imprecision in both first- and second-order probabilities. We have argued that when ideal first-order probabilities P_T are meaningful, they will often be imprecise. Even when P_T is precise, as when it is an unknown sampling model, imprecise second-order probabilities will usually be needed. It is unrealistic to suppose that, when first-order assessments yield only a large class \mathcal{M} of previsions on $\mathcal{L}(\Omega)$, a precise second-order model P_2 can be constructed on the more complicated space $\mathcal{L}(\mathcal{M})$.⁹

Consider, then, a class Θ of coherent lower previsions $\underline{P}(\cdot|\theta)$, each defined on $\mathcal{L}(\Omega)$, as hypotheses about the true probabilities. Suppose You assess a coherent lower prevision \underline{P}_2 on $\mathcal{L}(\Theta)$, representing Your beliefs about the true probabilities. From section 6.7, the natural extension of these models to first-order probabilities on $\mathcal{L}(\Omega)$ is $\underline{P}_1(X) = \underline{P}_2(X^*)$, where $X^* \in \mathcal{L}(\Theta)$ is the gamble defined by $X^*(\theta) = \underline{P}(X|\theta)$. As before, \underline{P}_1 represents Your current behavioural dispositions.

If the second-order probabilities \underline{P}_2 are non-vacuous, then there is some gamble X and $\theta \in \Theta$ such that $\underline{P}_1(X) > \underline{P}(X|\theta)$. This means that You are currently prepared to pay more than $\underline{P}(X|\theta)$ for X , so $\underline{P}(\cdot|\theta)$ cannot possibly be an exhaustive model for Your current dispositions. As before, it is inconsistent to interpret the hypotheses $\underline{P}(\cdot|\theta)$ as exhaustive models for current beliefs, unless Your beliefs about these hypotheses are vacuous. If \underline{P}_2 is vacuous, however, then $\underline{P}_1(X) = \inf \{\underline{P}(X|\theta) : \theta \in \Theta\}$, so that \underline{P}_1 is just the lower envelope of the hypotheses $\underline{P}(\cdot|\theta)$.

The second-order probabilities \underline{P}_2 can (coherently) be non-vacuous provided the models $\underline{P}(\cdot|\theta)$ are not regarded as hypotheses about current behavioural dispositions. Instead they might have an aleatory or logical interpretation. The type of natural extension outlined above can then

be used as an assessment strategy for constructing coherent first-order probabilities \underline{P}_1 from second-order beliefs about the correct model $\underline{P}(\cdot|\theta)$.¹⁰

5.10.6 Conclusion

Second-order probabilities are attractive to Bayesians because they enable the imprecise model \mathcal{M} to be reduced to a precise probability P_1 . There are several objections to theories of second-order probability which assume that both P_T and P_2 are precise. Fundamental objections concerning the meaning and precision of P_T were discussed in section 5.9. It will be complicated in general to construct a precise second-order distribution P_2 , especially when \mathcal{M} is a large or complicated set. In practice there is considerable arbitrariness in the choice of P_2 . Because of the arbitrariness and practical difficulties of assessment, Bayesians often take P_2 to be some kind of ‘noninformative’ distribution; see section 5.12.4 for simple examples. To remove the arbitrariness and to properly reflect the limitations of evidence, it seems necessary to allow imprecise second-order probabilities.

There are fewer objections to a theory of second-order probabilities which allows both P_T and P_2 to be imprecise, along the lines of section 5.10.5, although serious difficulties remain concerning the definition of ‘true’ probabilities, the assessment of P_2 and the complexity of computations. The imprecision in \mathcal{M} will be reduced, but not eliminated, by assessing an imprecise, non-vacuous, second-order model.

5.11 Fuzzy sets

There are many accounts of reasoning and decision making which are based on the theory of fuzzy sets.¹ In this section we discuss one of these approaches, known as **fuzzy decision analysis**, which can be regarded as another elaboration of Bayesian sensitivity analysis. The imprecision in probabilities and utilities is modelled, in this approach, through **membership functions** defined on the sets of possible probabilities and utilities.²

5.11.1 Membership functions

In ordinary language, there is considerable vagueness or ambiguity associated with terms such as ‘tall’. It may be unclear whether a specific person should be described as ‘tall’, because there is no precise height that separates tall from non-tall persons.³

The theory of fuzzy sets, introduced by Zadeh (1965), aims to model this ambiguity. Instead of classifying a person x as ‘tall’ or ‘not tall’, You

(subjectively) assign him a **degree of membership** $\mu_t(x)$, between zero and one, in the fuzzy set of tall people. The number $\mu_t(x)$ measures the ‘degree of possibility’ or ‘degree of truth’ that person x is tall. A person who is definitely tall will have $\mu_t(x) = 1$, one who is definitely not tall will have $\mu_t(x) = 0$, and intermediate values are used for the ambiguous cases. If there are no ambiguous cases then the membership function μ_t is simply the indicator function of the set of tall people (section 2.7.1). In general, membership functions are generalizations of indicator functions which allow degrees of membership in a set.

5.11.2 Fuzzy probability judgements

This idea can be applied to probability judgements by allowing degrees of membership in the set \mathcal{M} of ‘possible’ probabilities. Consider, for example, the judgement that an event A is probable. In section 4.4 we suggested that this be modelled by $P(A) = \frac{1}{2}$, which corresponds to the interval of precise probabilities $[\frac{1}{2}, 1]$ for A . The corresponding membership function is just the indicator function of this interval. However, the term ‘probable’ is somewhat ambiguous in ordinary language. For some people, the judgement ‘ A is probable’ might mean that the ‘true probability’ is more likely to be in the interval $(0.6, 0.7)$ than in $(0.9, 1)$, and that could be modelled by a membership function μ_p which assigned higher degrees of membership to precise probabilities in the first interval than in the second.⁴

Similarly Watson, Weiss and Donnell (1979) represent the judgement ‘pretty likely’ by a membership function which is roughly constant for probabilities between 0.65 and 0.9, and zero outside the interval $[0.55, 1]$. They translate the judgement ‘about 20%’ into a membership function that is zero outside $(0.1, 0.3)$, one at 0.2, linearly increasing on $(0.1, 0.2)$, and linearly decreasing on $(0.2, 0.3)$. These translations are quite arbitrary, especially as no clear interpretation of membership functions is suggested.⁵

More generally, suppose that, using the methods of Chapter 4, You assess a class $\mathcal{M} = \mathcal{M}(P)$ of linear previsions. You may be able to distinguish some linear previsions in \mathcal{M} as ‘more possible’ than others, and represent these judgements by a membership function μ_p , defined on the class \mathcal{P} of all linear previsions, with $\mu_p(P) = 0$ unless $P \in \mathcal{M}$.

This approach can be regarded as an elaboration or generalization of Bayesian sensitivity analysis, which corresponds to the special case where μ_p is the indicator function of \mathcal{M} . (Generally, \mathcal{M} is replaced by a fuzzy set.) The membership function μ_p represents some sort of second-order beliefs about the true probabilities P_T , and it is therefore comparable with the second-order probability distributions discussed in section 5.10.

5.11.3 Fuzzy decision analysis

The theory of decision making that is outlined by Watson *et al.* (1979) and Freeling (1980) adopts the framework of Bayesian decision theory, but allows fuzziness in probabilities and utilities. That leads to fuzziness in expected utilities and in preferences between actions. A fuzzy decision analysis proceeds in three steps:

1. Assess a membership function μ_p on the class of precise probabilities, and similarly assess a membership function μ_u on the class of linear utility functions.
2. For each feasible action a , compute the membership function μ_a for its expected utility using the rules given in section 5.11.6: $\mu_a(z)$ is the maximum value of $\min\{\mu_p(P), \mu_u(U)\}$, over all combinations of precise probabilities P and utility functions U which give expected utility z .
3. Compare the membership functions μ_a for the various actions a . Two actions a and b can be compared, for example, by computing the membership function for their difference in expected utility.⁶

In the rest of this section we discuss some fundamental difficulties in applying this approach.

5.11.4 Imprecision versus ambiguity

Fuzzy sets have been advocated as a way of ‘handling imprecision’ in probabilities and utilities. However, fuzzy sets were introduced for the rather different purpose of modelling the ambiguity of ordinary language. Compare the statements ‘he is fairly tall’ and ‘ $P(A)$ is about 0.2’, which are both imprecise and ambiguous, with ‘he is more than 1.8 metres tall’ and ‘ $P(A)$ is between 0.15 and 0.25’, which are imprecise but unambiguous. Imprecise judgements can be expressed without ambiguity, and they can be modelled without fuzzy sets.⁷

Nor is it clear that the ambiguity of ordinary language is best handled through fuzzy sets. Imprecision in probability judgements, insofar as it reflects real limitations in evidence, is desirable. Ambiguity is undesirable, however, and one might try to eliminate it during elicitation by clarifying the definition of ambiguous terms in order to determine the behavioural meaning of each probability judgement. It seems feasible to replace ambiguous judgements like ‘ $P(A)$ is about 0.2’ with unambiguous behavioural judgements like ‘I would accept either of the gambles $A - 0.15$ or $0.25 - A'$.

A related problem is that, whereas there is a clear correspondence between the behavioural judgements and their imprecise probability models, the mathematical precision of a membership function μ_p (defined on the

probability interval $[0, 1]$) seems quite inappropriate to model a vague judgement like ' $P(A)$ is about 0.2'. There is a mismatch between the vagueness of ordinary language and the precision of its mathematical representation.

5.11.5 Meaning and assessment of μ

What does it mean to say that x has degree of membership 0.3 in the class of tall people, or that the precise probability $P(A) = 0.1$ has degree of membership 0.8 in the class of possible probabilities? Little has been done to explain the meaning of these numbers. According to Watson *et al.* (1979), ' $\mu_p(p)$ is the degree to which p belongs to the set of possible values for this probability'.⁸ It appears that μ_p represents subjective, second-order judgements concerning the 'true' value of a precise probability, but beyond that the interpretation of the 'degree of possibility' $\mu_p(p)$ remains unclear.

People can be ordered according to height, and this ordering should be reflected in the membership function μ_t for the term 'tall'. If $\mu_t(x) > \mu_t(y)$, then we may infer that x is taller than y . But it is unclear what it means to say that one probability value p_1 has a higher degree of membership than another value p_2 . Does this mean that p_1 is *more likely* than p_2 to be the true probability p_T ?

In fact, it is most natural to interpret the membership function μ_p as some kind of second-order probability. For example, μ_p might be a multiple of the second-order probability density function for the true probability P_T ; or $\mu_p(P)$ might be interpreted as the second-order probability that P belongs to the set \mathcal{M} . Such probabilistic interpretations are untenable, however, because they do not support the standard rules for combining membership functions (section 5.11.6), and they would anyway reduce this approach to the theory of second-order probabilities.⁹

The absence of an interpretation of μ_p , behavioural or otherwise, is a major problem. Because it is not clear what μ_p means, it is not clear how to assess it, how to interpret the conclusions of a fuzzy analysis (which are also expressed in terms of membership functions), or how to justify the rules of combination for membership functions.

A further difficulty in assessment is that a fuzzy analysis requires considerably more input than a standard decision analysis or sensitivity analysis. After eliciting the class of linear previsions \mathcal{M} , which may already be a complicated set, You must define a precise membership function μ_p on \mathcal{M} . That will be very difficult to do in general.¹⁰

One might try to alleviate the difficulties of assessing μ_p by using standard membership functions such as those suggested in section 5.11.2, despite their arbitrariness, to model judgements expressed in ordinary language. But that seems unwise, as the meaning of ordinary language is both personal and

context-dependent. My usage of 'tall' differs from yours, 'tall' applies differently to people and buildings, and whether a person is 'tall' depends on whether he is in China or the USA. Similarly the judgement 'about 20%' may have different precision in different contexts. The membership function μ_p will presumably vary according to who makes the judgement, and in what context.¹¹

5.11.6 Fuzzy calculus

The three basic rules for combining membership functions were suggested by Zadeh (1965). These rules define the membership functions for the fuzzy sets A^c , $A \cap B$ and $A \cup B$ in terms of the membership functions for A and B , by $\mu_{A^c}(x) = 1 - \mu_A(x)$, $\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$, and $\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}$.

The simplicity of these rules seems to be one reason for the popularity of fuzzy sets. The rules need some more compelling justification, however, in view of their fundamental role in fuzzy reasoning, and this would require a more definite interpretation of membership functions than has so far been established.¹²

It is implausible that a single rule of combination will suffice to determine the membership function of $A \cap B$, irrespective of the relationship between properties A and B . That is clear if $\mu_A(x)$ is interpreted as the probability that x has property A . Then Zadeh's rules for determining $\mu_{A \cap B}$ and $\mu_{A \cup B}$ would be appropriate when A and B are fully dependent properties (i.e., one property implies the other), but the different rules

$$\mu_{A \cap B}(x) = \mu_A(x)\mu_B(x) \quad \text{and} \quad \mu_{A \cup B}(x) = 1 - (1 - \mu_A(x))(1 - \mu_B(x))$$

would be appropriate when A and B are independent properties (i.e., the probability that x has one property does not depend on whether it has the other property). A continuum of other rules would be needed for the intermediate cases.¹³

This probabilistic interpretation is obviously incompatible with Zadeh's calculus, but it is difficult to imagine an alternative interpretation of μ under which the degree of dependence between properties A and B is irrelevant to the calculation of $\mu_{A \cap B}$. The fuzzy calculus may be attractive to users because it does not require them to assess such degrees of dependence, but it seems unrealistic to expect to reach sensible conclusions without these assessments.

Zadeh's combination rules are used in fuzzy decision analysis to compute the membership function for expected utility (section 5.11.3). It seems especially inappropriate to combine the membership functions for probability and utility using the minimum operator, as (intuitively) there will

rarely be any strong degree of dependence (in second-order beliefs) between probabilities and utilities.

5.11.7 Conclusion

If fuzzy sets have a useful role to play, it is in modelling the ambiguity of ordinary language. But ambiguity is only one potential source of imprecision in probabilities and, unlike some other sources of imprecision, it can be eliminated through careful elicitation. The membership functions μ_p that are chosen to model ambiguous probability judgements seem both arbitrary and inappropriately precise. No clear interpretation of μ_p has been established. Its assessment requires substantial input from a user, in addition to that needed for sensitivity analysis. (The extra assessments are comparable to those needed to determine second-order probabilities.)

Fuzzy decision analysis may be seen as an elaboration of Bayesian sensitivity analysis, but it does not appear to add anything useful to sensitivity analysis. On the contrary, fuzzy analysis may obscure the decision problem, by adding second-order structure which is difficult to assess and whose meaning is unclear.¹⁴

5.12 Maximum entropy

The maximum entropy approach can be viewed as another extension of sensitivity analysis, which selects a unique linear prevision from the class \mathcal{M} of ‘possible’ previsions. The approach is based on the **principle of maximum entropy**, proposed by Jaynes (1957, 1968), which states that You should select the linear prevision from \mathcal{M} that maximizes entropy. Most of this section is concerned with evaluating the arguments for and against the principle of maximum entropy.¹

Suppose that the possibility space of interest, Ω , is finite. The **entropy** of an additive probability measure on Ω is defined to be

$$H_P = - \sum_{\omega \in \Omega} P(\omega) \log P(\omega),$$

where $P(\omega) \log P(\omega)$ is taken to be zero when $P(\omega) = 0$. The entropy H_P is interpreted as a measure of the amount of uncertainty in the probability distribution P . The maximum entropy approach is based on the following simple principle.

5.12.1 Principle of maximum entropy (PME)

When Your partial information concerning Ω determines a class \mathcal{M} of precise probabilities, You should select, and act according to, the precise probability measure P_0 which maximizes the entropy H_P over P in \mathcal{M} .²

According to Jaynes (1957, 1968) and Rosenkrantz (1977), the PME is a basic principle of rationality, and P_0 is the correct logical probability measure to adopt when Your partial information determines only the class \mathcal{M} .

Although its scope is more general, Jaynes (1968) and Tribus (1969) advocate the PME primarily as a way of selecting Bayesian prior probabilities in problems where there is little prior information. It is just these problems where the choice of prior probabilities is most arbitrary and the Bayesian approach is most controversial.³

5.12.2 Examples of the PME⁴

The simplest application of the PME is in the case where there is no prior information, so that $\mathcal{M} = \mathcal{P}$, the class of all linear previsions. The PME selects the uniform distribution on Ω , which maximizes entropy amongst all distributions, to represent complete ignorance. The classical principle of indifference (section 5.5.1) can be regarded as a special case of the PME.

As a second application, suppose You assess the previsions of the gambles X_j to be precisely μ_j ($1 \leq j \leq k$), so that \mathcal{M} consists of all linear previsions P satisfying the k constraints $P(X_j) = \mu_j$. (Assume that these linear constraints are independent.) It can be shown, by a standard optimization using Lagrange multipliers, that the maximum entropy distribution is $P_0(\omega) = \lambda_0 \exp(-\sum_{j=1}^k \lambda_j X_j(\omega))$, where the constants $\lambda_0, \lambda_1, \dots, \lambda_k$ are determined by the k linear constraints and the further constraint $\sum_{\omega \in \Omega} P_0(\omega) = 1$.

5.12.3 Arguments for the PME

Next we outline the arguments presented in favour of the PME, and evaluate their strength. A common argument for the PME is that, because entropy is a measure of uncertainty or ‘noninformativeness’, the distribution P_0 with maximum entropy is ‘maximally noncommittal with regard to missing information’, or ‘as noninformative as possible’, or ‘minimally prejudiced’ amongst all precise probability distributions.⁵ To defend the PME, it is necessary to answer the following three questions:

1. Why should any precise probability model P_0 be adopted?
2. Why choose P_0 to maximize uncertainty, or minimize information, within \mathcal{M} ?
3. Why is entropy the appropriate measure of uncertainty or information?

(a) Need for precision

The basic motivation for the PME seems to be that Bayesian’s need to obtain a precise prior distribution P_0 even when their prior information yields only

an imprecise class \mathcal{M} . The PME is a simple rule for generating a precise P_0 , and Bayesians can apply it without making the difficult assessments of second-order probabilities that are needed for the method discussed in section 5.10. If one needs to select a precise P_0 from \mathcal{M} , then the PME has some appeal.⁶

Our view, of course, is that any choice of precise P_0 will misrepresent the limited prior information. The solution is to accept \mathcal{M} itself as a model for beliefs, and to recognize the indeterminacy and indecision that entails.

(b) Axiomatic derivations of entropy

Entropy is a basic concept in Shannon's theory of information, primarily because of its role in his coding theorem. In addition, many axiom systems have been proposed that lead to entropy as the unique measure of 'information'.⁷ It has been established that entropy measures one important sense of 'information': the average (or expected) number of binary symbols needed to specify a random state ω , when each ω occurs with known aleatory probability. Although this is of great importance in the theory of communications, it does not seem to be the kind of 'information' that is most important in statistical inference.⁸

The various axiomatic derivations of entropy proceed by formalizing some intuitively desirable properties of a measure of information, such as additivity across independent sources of information, and showing that entropy is the unique measure which has all these properties. It is assumed that the measure is a functional defined on precise probability distributions P , which means that the 'amount of information' on which P is based can be recovered from P alone. This assumption seems to be mistaken, in view of the paradox of ideal evidence and the other difficulties noted in section 5.3.

(c) Success in generating empirical hypotheses

Maximum entropy ideas have been remarkably successful in the fields of statistical mechanics and classical thermodynamics. In these fields the PME, together with knowledge of physical constraints, has been used to generate empirical hypotheses concerning aleatory probabilities, and the hypotheses have been experimentally verified.⁹

We have no objection to this use of the PME as a source of empirical hypotheses.¹⁰ But its success in predicting the correct aleatory probabilities in some physical problems does not justify use of the PME as a general method for determining rational epistemic (logical) probabilities. In typical applications of the PME, there are no underlying aleatory probabilities.

(d) Agreement with relative frequencies

Suppose we are able to carry out n repetitions of an experiment with

outcomes in Ω , and obtain relative frequency distribution r_n over Ω . It can be shown that, for sufficiently large n , a great majority of those outcome sequences for which r_n is in \mathcal{M} will have r_n arbitrarily close to P_0 , the distribution which maximizes entropy in \mathcal{M} . Thus 'the distribution predicted by maximum entropy can be realized experimentally in overwhelmingly more ways than can any other'.¹¹

This result is combinatorial rather than probabilistic. If we restrict attention to outcome sequences for which r_n is in \mathcal{M} , the result tells us that for *most* such sequences r_n is close to P_0 , but it does not tell us that we are *likely* to obtain r_n close to P_0 . We could reach the latter conclusion only by adding probabilistic assumptions, e.g. that the outcomes are equally probable and independent, which are quite unwarranted in general. There is no reason to expect that something close to the maximum entropy distribution will be 'realized experimentally'.¹²

(e) Logarithmic scoring rules

In our view, none of the preceding arguments for the PME has much force. We consider finally a decision-theoretic argument which shows that, in certain specific decision problems, it is at least reasonable to act according to the PME. The idea is to elicit probabilities through a scoring rule which effectively forces You to select a single probability measure P_0 from the class \mathcal{M} . If the scoring rule has a logarithmic form, then the response P_0 which maximizes entropy within \mathcal{M} is the **minimax** decision.

Formally, You are required to specify a probability mass function Q on the finite space Ω . You will receive the uncertain reward $\log Q(\omega)$. How should You choose Q ? Assuming that Your beliefs are represented by the lower prevision \underline{P} , the maximal choices Q are just those corresponding to linear previsions in $\mathcal{M} = \mathcal{M}(\underline{P})$. Any such choice is reasonable.

One way to choose a unique Q from \mathcal{M} is to maximize the satisfaction level $\underline{P}(\log Q)$. This is the \underline{P} -minimax decision rule, discussed in sections 3.9.7 and 5.6.6. The minimax theorem (Appendix E6) can be used to show that

$$\begin{aligned} \max_{Q \in \mathcal{P}} \underline{P}(\log Q) &= \max_{Q \in \mathcal{P}} \min_{P \in \mathcal{M}} P(\log Q) = \min_{P \in \mathcal{M}} \max_{Q \in \mathcal{P}} P(\log Q) \\ &= \min_{P \in \mathcal{M}} P(\log P) = -\max_{P \in \mathcal{M}} H_P, \end{aligned}$$

where $H_P = -P(\log P)$ is the entropy of P .¹³ It follows that $\underline{P}(\log Q)$ is maximized over probability mass functions Q by the unique Q which maximizes entropy within \mathcal{M} . So the minimax rule is to select the maximum entropy distribution from \mathcal{M} .¹⁴

This provides some support for the PME in a very special kind of decision problem, but not as a general rule for forming beliefs. Other scoring rules

will lead, by the same argument, to different choices Q . Consider, for example, the quadratic scoring rule $S(Q) = 2Q - \sum_{\omega \in \Omega} Q(\omega)^2$.¹⁵ The minimax theorem gives

$$\max_{Q \in \mathcal{P}} P(S(Q)) = \min_{P \in \mathcal{M}} P(S(P)) = \min_{P \in \mathcal{M}} \sum_{\omega \in \Omega} P(\omega)^2.$$

It follows that the minimax decision is to select the unique Q from \mathcal{M} which minimizes the sum of squared probabilities $\sum_{\omega \in \Omega} Q(\omega)^2$. In general, this Q will not maximize entropy over \mathcal{M} , so the quadratic and logarithmic scoring rules will lead to different choices of Q .¹⁶

5.12.4 Other objections to the PME

(a) Complete ignorance

Since the PME yields the uniform distribution as a representation of complete ignorance, the objections raised in section 5.5.1 apply also to the PME. In particular, the probability assigned to an event A by the PME depends on the possibility space Ω in which A is embedded.

(b) Conditioning on new information

The inferences produced by the PME may depend on whether it is applied before or after new information is obtained. To see that, let $\Omega = \{W, L, D\}$ represent the possible outcomes (win, lose or draw) of a football game. Suppose that You obtain two items of information about the game. Before the game, You learn that the home team has won more than half of previous games between the teams. On that basis You judge that W is probable, modelled by the class of linear previsions $\mathcal{M}_1 = \{P \in \mathcal{P}: P(W) \geq \frac{1}{2}\}$. After the game, You learn a second piece of information, that the game was not drawn. The class \mathcal{M}_1 should then be updated by conditioning on the event $\{W, L\}$, using the generalized Bayes' rule (Theorem 6.4.1).¹⁷ This yields the new class $\mathcal{M}_2 = \{P \in \mathcal{P}: P(W) \geq \frac{1}{2}, P(D) = 0\}$, which models beliefs about the outcome based on both items of information.

If the PME is applied before the game, to \mathcal{M}_1 , it yields the probability mass function $Q_1 = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, which is then adopted as Your precise probability model. When You learn that the game was not drawn, You should condition Q_1 on the event $\{W, L\}$, using Bayes' rule, yielding the final mass function $P_1 = (\frac{2}{3}, \frac{1}{3}, 0)$.

Alternatively, You could apply the PME to the class \mathcal{M}_2 , after both items of information are received. But this yields a different mass function $P_2 = (\frac{1}{2}, \frac{1}{2}, 0)$. So the PME can produce different inferences from the same information, depending on when it is applied in processing the information.

This example also illustrates that the PME will often give very different answers from the method of assessing second-order probabilities on \mathcal{M} .

(section 5.10). The latter will typically generate a composite first-order distribution in the relative interior of \mathcal{M} , whereas the maximum entropy distribution will often be on the boundary of \mathcal{M} . In the football example, a uniform second-order distribution over \mathcal{M}_1 yields $Q_3 = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ as the composite first-order mass function, which gives $P_3 = (\frac{4}{5}, \frac{1}{5}, 0)$ when conditioned on $\{W, L\}$. Alternatively, a uniform second-order distribution on \mathcal{M}_2 generates $P_4 = (\frac{3}{4}, \frac{1}{4}, 0)$ as the final first-order mass function.

The four posterior distributions P_1 , P_2 , P_3 and P_4 assign quite different probabilities to 'win' ($\frac{2}{3}$, $\frac{1}{2}$, $\frac{4}{5}$ and $\frac{3}{4}$ respectively). Is one of these the 'correct' probability? We think not. There is evidently some arbitrariness in the choice of a precise probability distribution from \mathcal{M}_2 , and the only realistic model for the limited information seems to be \mathcal{M}_2 itself.

(c) Infinite entropy

When Ω is countably infinite, there are many countably additive distributions with infinite entropy.¹⁸ In this case the implications of the PME are unclear.

(d) Continuous spaces

Jaynes (1968) defines the entropy of a probability density function f on a continuous space Ω to be

$$H_f = - \int f(\omega) \log(f(\omega)/h(\omega)) d\omega,$$

where h is the density function of a suitable 'invariant measure' which is supposed to model complete ignorance about Ω . The density f is then chosen, subject to any constraints, to maximize H_f . Thus the choice of f depends on finding a suitable measure to represent complete ignorance, which is subject to the difficulties discussed in section 5.5.¹⁹

When ω is a real location parameter, for example, Jaynes advocates the improper uniform density as the appropriate density h . In addition to the previous objections (section 5.5.4), there is the new difficulty that the entropy may be unbounded or infinite. When there are no constraints on f , for example, or when f has specified quantiles, there are many densities with infinite entropy, and the PME does not yield a unique density. In statistical problems, parameter spaces are typically continuous and these difficulties seem to be quite common.

5.12.5 Conclusion

None of the arguments for the PME, when regarded as a general method for generating precise epistemic probabilities, is at all compelling.²⁰ (We accept that the PME can be a useful method of generating aleatory hypotheses.) Since the PME is a generalization of the classical principle of

indifference, and gives rise in special cases to the noninformative priors criticized in section 5.5, it is subject to the earlier objections.

The main motivation for the PME is that Bayesians need to select precise probabilities P_0 even when their information determines only an imprecise class \mathcal{M} . Selecting P_0 to maximize entropy may be as reasonable as other methods of selection, such as minimizing the sum of squared probabilities or assigning a second-order distribution on \mathcal{M} . But any such method is somewhat arbitrary, as it is creating precision out of ignorance. The precision is unwarranted.

5.13 The Dempster–Shafer theory of belief functions

To end this chapter, we outline the theory of belief functions that has been developed by Dempster and Shafer.¹ A belief function is a special type of coherent lower probability. The theory is especially interesting because it emphasizes the process of constructing probabilities from evidence and, unlike the approaches described in sections 5.9–5.12, it does not rely on a sensitivity analysis interpretation of upper and lower probabilities. The theory uses Dempster's rule to combine belief functions based on separate bodies of evidence. One of our aims is to clarify the independence conditions under which Dempster's rule is applicable, and to show how restrictive these conditions are. In some cases, Dempster's rule produces a sure loss.

5.13.1 Belief functions

Mathematically, belief functions are coherent lower probabilities that satisfy an extra property of complete monotonicity, defined as follows.² Suppose that \mathcal{A} is a field of subsets of Ω and \underline{P} is a lower probability defined on \mathcal{A} , which satisfies $\underline{P}(\emptyset) = 0$, $\underline{P}(\Omega) = 1$ and $\underline{P}(A) \geq 0$ for all $A \in \mathcal{A}$. Then \underline{P} is called **completely monotone** if, for any events A_1, A_2, \dots, A_n in \mathcal{A} and $n \geq 2$,

$$\underline{P}\left(\bigcup_{i=1}^n A_i\right) \geq \sum_{J \subset \{1, 2, \dots, n\}} (-1)^{|J|+1} \underline{P}\left(\bigcap_{i \in J} A_i\right),$$

where $|J|$ denotes the cardinality of the set J . All belief functions (completely monotone lower probabilities) are coherent.³

We will restrict attention to the case of finite Ω . This case has been studied in detail by Shafer (1976), who gives a more useful characterization of belief functions in terms of probability mass functions. A lower probability \underline{P} (defined on all subsets of Ω) is a **belief function** if and only if it can be written in the form $\underline{P}(A) = \sum_{B \subseteq A} m(B)$ for all sets A , where m is a probability mass function defined on all subsets of Ω , that is, $m(\emptyset) = 0$, $m(B) \geq 0$ for all subsets B , and $\sum_{B \subseteq \Omega} m(B) = 1$. The function m is called the **probability assignment**. It is helpful to think of each probability mass $m(B)$ as a fluid mass that is restricted to the set B but is free to move to any of the elements

of B . Unless B is a subset of A , the mass $m(B)$ is free to move outside A . So the lower probability $\underline{P}(A) = \sum_{B \subseteq A} m(B)$ is the minimum probability mass that must lie in A , or the probability that is *committed* to A . Similarly, the conjugate upper probability $\bar{P}(A) = 1 - \underline{P}(A^c) = \sum_{B \cap A \neq \emptyset} m(B)$ is the maximum probability mass that can move to A .

The function m is completely determined by \underline{P} through the Möbius inversion formula

$$m(B) = \sum_{A \subseteq B} (-1)^{|B-A|} \underline{P}(A),$$

and \underline{P} is a belief function if and only if the function m defined by this formula is a probability mass function. It is usually most convenient to specify a belief function \underline{P} through its probability assignment m .⁴

The class of belief functions includes all additive probabilities (characterized by $m(B) = 0$ unless $|B| = 1$), and some non-additive models defined in section 2.9, including the vacuous lower probability (which has $m(\Omega) = 1$), all linear–vacuous mixtures, and all zero–one valued coherent lower probabilities. There are many coherent lower probabilities which are not belief functions; a simple example is given in section 5.13.4.

5.13.2 Multivalued mappings

In the theory of Dempster (1967a), belief functions are generated by the multivalued mappings discussed in section 4.3.5. Suppose that the possibility space of interest, Ω , is related to an underlying space $\Psi = \{\psi_1, \psi_2, \dots, \psi_n\}$ through a multivalued mapping A , so that each state ψ_i is consistent with all states in the non-empty subset $A(\psi_i)$ of Ω . Suppose that Your beliefs about Ψ are modelled by a precise probability measure P . (For example, P might describe a known stochastic mechanism which generates ψ .) If You have no additional information concerning Ω then Your beliefs about it can be represented by the probability assignment m that is induced by P through the multivalued mapping. This is defined by transferring the probability $P(\psi_i)$ to the subset $A(\psi_i)$, so that $m(B) = P(\{\psi_i : A(\psi_i) = B\})$. We will assume, for simplicity, that the sets $A(\psi_i)$ are distinct,⁵ so m is defined by $m(A(\psi_i)) = P(\psi_i)$ for $1 \leq i \leq n$, $m(B) = 0$ for other sets B .

Shafer (1981a, 1982a) has developed the idea of a multivalued mapping by putting more emphasis on the evidence from which belief functions are constructed. Shafer appears to regard the underlying states ψ_i as different ways of explaining or interpreting the evidence x that You have in a particular problem.⁶ In some problems, the states ψ_i can be identified with hypothetical mechanisms which might have produced the evidence x . On this interpretation, the model rests on the following three assumptions:

1. $\Psi = \{\psi_1, \dots, \psi_n\}$ is an exhaustive set of possible states or mechanisms which might have produced the evidence x .

2. Assuming that the underlying state is ψ_i , what You would learn about Ω from x is just that ω belongs to a non-empty subset $A_i = A^x(\psi_i)$ of Ω . (Here A^x is a multivalued mapping.)
3. After observing x , You assign precise posterior probabilities $p_i = P(\psi_i|x)$ to the hypotheses ψ_i .

Under these assumptions, Your beliefs about Ω after observing the evidence x can be represented by the probability assignment m^x defined by $m^x(A_i) = p_i$, or by its corresponding belief function \underline{P}^x .⁷

5.13.3 Example: an unreliable witness

One of the simplest applications of belief functions is in modelling unreliable observations. Suppose that Your evidence x is a report from an unreliable witness that event C has occurred.⁸ You might consider two possible explanations: either the witness really observed C , or he observed nothing at all. These hypotheses are represented by ψ_1 and ψ_2 , with multivalued mapping $A^x(\psi_1) = C$ and $A^x(\psi_2) = \Omega$. If You assess the posterior probability that he did observe C (conditional on the report that he did) to be precisely $P(\psi_1|x) = p_1$, then You obtain the probability assignment $m^x(C) = p_1$, $m^x(\Omega) = 1 - p_1$. This corresponds to the belief function defined by $\underline{P}^x(B) = 0$ unless $B \supset C$, $\underline{P}^x(B) = p_1$ if $B \supset C$ but $B \neq \Omega$, and $\underline{P}^x(\Omega) = 1$. Thus You are not certain that C has occurred, since $\underline{P}^x(C) = p_1 < 1$, but You are not willing to bet against C , since $\bar{P}^x(C) = 1 - \underline{P}^x(C^c) = 1$.

The assessment strategy suggested by Dempster and Shafer, which involves assessing precise probabilities for alternative explanations of the evidence that are related to Ω through a multivalued mapping, is a useful source of coherent lower probabilities.⁹ It can be especially useful in modelling testimony or observations that are not completely reliable, as illustrated in the preceding example.

Belief functions can be accommodated in our theory as a special type of coherent lower probability.¹⁰ But the class of belief functions is too small for a theory of imprecise probabilities; there are many reasonable states of uncertainty that cannot be modelled by any belief function. That can be illustrated by simple examples such as the following.

5.13.4 Example: two tosses of a coin

Suppose that a fair coin is tossed twice, in such a way that the outcome of the second toss may depend on the outcome of the first, but You are completely ignorant about the degree of dependence. Let H_1 , T_1 , H_2 , T_2 denote the possible outcomes of the two tosses. Since the coin is known to be fair, we require upper and lower probabilities \bar{P} and \underline{P} that satisfy

- (a) $\underline{P}(H_1) = \bar{P}(H_1) = \frac{1}{2}$, and (b) $\underline{P}(H_2) = \bar{P}(H_2) = \frac{1}{2}$. To model ignorance about the degree of dependence we require also (c) $\underline{P}(H_1 \cap H_2) = 0$ and $\bar{P}(H_1 \cap H_2) = \bar{P}(H_1) = \frac{1}{2}$, since the occurrence of H_1 may never lead to H_2 , or it may always do so.

It is easy to verify that there is a unique coherent lower probability \underline{P} which satisfies these constraints. It is the lower envelope of all additive probability measures P which assign $P(H_1) = P(H_2) = \frac{1}{2}$ and have an arbitrary degree of dependence between tosses.¹¹ To see that \underline{P} is not a belief function, note that $\underline{P}(H_1) = \underline{P}(H_2) = \frac{1}{2}$, $\underline{P}(H_1 \cap H_2) = 0$, and $\underline{P}(H_1 \cup H_2) \leq \bar{P}(H_1) + \bar{P}(H_2) - \bar{P}(H_1 \cap H_2) = \frac{1}{2}$, using property 2.7.4(h). Hence $\underline{P}(H_1 \cup H_2) < \underline{P}(H_1) + \underline{P}(H_2) - \underline{P}(H_1 \cap H_2)$, which violates the condition of complete monotonicity with $n = 2$.¹²

5.13.5 Dempster's rule of combination

The Dempster–Shafer theory relies heavily on Dempster's rule of combination. According to Shafer (1981a), ‘Dempster's rule of combination is the most important single tool of the theory’. The general strategy is to break up Your total evidence into simpler, unrelated pieces, then construct belief functions from each piece of evidence, and finally combine the belief functions by Dempster's rule.

Suppose that evidence x is regarded as arising from one of the states in $\Psi = \{\psi_1, \dots, \psi_r\}$, related to Ω through the multivalued mapping $A^x(\psi_i) = A_i$, while a separate body of evidence y arises from $\Phi = \{\phi_1, \dots, \phi_k\}$, related to Ω through the multivalued mapping $A^y(\phi_j) = B_j$. Suppose also that You assess precise probabilities $P(\psi_i|x) = p_i$ and $P(\phi_j|y) = q_j$, so that the probability assignment based on x is $m^x(A_i) = p_i$ ($1 \leq i \leq n$), and that based on y is $m^y(B_j) = q_j$ ($1 \leq j \leq k$). Dempster's rule of combination is used to combine m^x , which represents Your beliefs about Ω based on evidence x alone, with m^y , which represents beliefs based on y alone, to give $m^{x,y}$, which models beliefs based on the combined evidence (x, y) .

Dempster's rule can be motivated by considering the joint states (ψ_i, ϕ_j) from which the combined evidence might have arisen. Assuming the joint state is (ψ_i, ϕ_j) , You know from the two multivalued mappings that $\omega \in A_i$ and $\omega \in B_j$, so $\omega \in A_i \cap B_j$. Thus the product space $\Psi \times \Phi$ of joint states is related to Ω through the multivalued mapping $A^{x,y}(\psi_i, \phi_j) = A_i \cap B_j$. In order to use this multivalued mapping it is necessary to assess the posterior probabilities of joint states, $P(\psi_i, \phi_j|x, y)$.

In general, some of the pairs (ψ_i, ϕ_j) may be impossible in the light of (x, y) , because the corresponding subsets $A_i \cap B_j$ are empty, but to simplify the discussion let us assume that all the sets $A_i \cap B_j$ are non-empty. Then You might judge that the two states are independent conditional on the

evidence, in the sense that the joint probabilities factorize as

$$P(\psi_i, \phi_j | x, y) = P(\psi_i | x)P(\phi_j | y) = p_i q_j.$$

If this independence judgement is made, then the multivalued mapping $A^{x,y}$ generates the probability assignment

$$m^{x,y}(C) = \sum_{A_i \cap B_j = C} p_i q_j = \sum_{A_i \cap B_j = C} m^x(A_i) m^y(B_j),$$

which represents beliefs based on the combined evidence. This rule for determining $m^{x,y}$ from m^x and m^y is a special case of **Dempster's rule of combination**.

5.13.6 Example: concordant witnesses

As a simple illustration of Dempster's rule, suppose there are two unreliable witnesses of the kind described in Example 5.13.3, who independently report the same observation C . Here ψ_1, ψ_2 denote the hypotheses that the report (x) of the first witness is reliable or unreliable, and ϕ_1, ϕ_2 denote that the second report (y) is reliable or unreliable. The two multivalued mappings are $A^x(\psi_1) = C$, $A^x(\psi_2) = \Omega$, and $A^y(\phi_1) = C$, $A^y(\phi_2) = \Omega$. The joint multivalued mapping is therefore $A^{x,y}(\psi_i, \phi_j) = C$ unless $i=j=2$, $A^{x,y}(\psi_2, \phi_2) = \Omega$.

Suppose You assess the reliability of each witness to be $m^x(C) = m^y(C) = 0.5$. Dempster's rule of combination then gives $m^{x,y}(C) = 0.75$ and $m^{x,y}(\Omega) = 0.25$, which corresponds to the independence judgement $P(\psi_2, \phi_2 | x, y) = P(\psi_2 | x)P(\phi_2 | y) = 0.25$. The corresponding belief function is $\underline{P}^{x,y}(C) = 0.75$ and $\bar{P}^{x,y}(C) = 1$. Corroboration of the first report by the second witness increases the lower probability $\underline{P}(C)$ from 0.5 to 0.75.¹³

5.13.7 Conditions for applying Dempster's rule¹⁴

When should Dempster's rule be used to combine evidence? Assuming the model given in section 5.13.5, it is reasonable to use Dempster's rule provided that it is reasonable to judge the underlying states conditionally independent, in the sense that $P(\psi_i, \phi_j | x, y) = P(\psi_i | x)P(\phi_j | y)$ for all values of i and j .¹⁵ That is equivalent to the conditions $P(\psi_i | x, y, \phi_j) = P(\psi_i | x)$ and $P(\phi_j | x, y, \psi_i) = P(\phi_j | y)$ for $1 \leq i \leq n$ and $1 \leq j \leq k$, which mean that, given one piece of evidence, Your beliefs about the state which produced it would be unchanged if You also learned the other piece of evidence and the state which produced it. It is this specific type of conditional independence that allows us to decompose the evidence (x, y) into the two separate pieces. Loosely, You must judge that the two sources of evidence are 'completely unrelated'.

Some insight into the strength of these conditions can be gained by considering Example 5.13.6. The judgements $P(\psi_2 | x, y, \phi_2) = P(\psi_2 | x) = 0.5$ and $P(\phi_2 | x, y, \psi_2) = P(\phi_2 | y) = 0.5$ are reasonable in the example, provided the two witnesses are 'independent' in the sense that learning that one made a mistake would not change beliefs about the reliability of the other.¹⁶ But the apparently similar condition $P(\psi_1 | x, y, \phi_1) = P(\psi_1 | x) = 0.5$ is not reasonable here. The extra information (y, ϕ_1) that the second witness really did observe C , which implies that C actually occurred, should increase Your probability that the first witness also observed it.

Writing $\alpha = P(\psi_1 | x, y, \phi_1)$ and using the first independence judgements, we find that $P(\psi_2, \phi_2 | x, y) = (1 - \alpha)/(3 - 2\alpha)$. This value generates $\underline{P}^{x,y}(C) = (2 - \alpha)/(3 - 2\alpha)$, which is greater than 0.75 (the value given by Dempster's rule) whenever $\alpha > 0.5$. If the probability α is assumed to be precise then the answer given by Dempster's rule, which corresponds to the assessment $\alpha = 0.5$, seems to be unreasonable.¹⁷

This does not settle the issue, however, because no precise assessment of α may be justified. If You make the minimal judgement that ψ_1 is *probable* conditional on the information (x, y, ϕ_1) , so that $\underline{P}(\psi_1 | x, y, \phi_1) = 0.5$ and $\bar{P}(\psi_1 | x, y, \phi_1) = 1$, then we obtain $\underline{P}(\psi_2, \phi_2 | x, y) = 0$ and $\bar{P}(\psi_2, \phi_2 | x, y) = 0.25$. From these imprecise assessments, the general model for multivalued mappings (section 4.3.5) generates $\underline{P}^{x,y}(C) = 0.75$ and $\bar{P}^{x,y}(C) = 1$, the same answer as Dempster's rule. This illustrates that Dempster's rule can give sensible answers in some problems where You are unable to assess precise probabilities for the joint states (ψ, ϕ) , so that the independence conditions used to derive the rule do not hold.

It seems that the precise independence conditions $P(\psi_i | x, y, \phi_j) = P(\psi_i | x)$ will rarely be satisfied in practice, because of the linkage suggested in the example. The two sources of evidence are linked through their multivalued mappings to the same space Ω , and knowledge of the state ϕ will typically provide information about ω and hence about the other state ψ .¹⁸ That should make us very cautious about using Dempster's rule, except when it is justified through explicit probability judgements. It should not be regarded as a substitute for such judgements.¹⁹

5.13.8 General version of Dempster's rule

An extreme case of the linkage between ψ and ϕ occurs when some of the sets $A_i \cap B_j = A^x(\psi_i) \cap A^y(\phi_j)$ are empty. In that case, knowledge of ϕ_j clearly does provide information about ψ : it tells You that ψ_i is not possible. You can no longer regard the two states as 'unrelated' or 'independent'.

Let Ξ denote the set of joint states that are possible in the light of the combined evidence, so $\Xi = \{(\psi_i, \phi_j) : A_i \cap B_j \text{ is non-empty}\}$. In general, Ξ

may be smaller than $\Psi \times \Phi$. Clearly You should assign $P(\psi_i, \phi_j | x, y) = 0$ unless $(\psi_i, \phi_j) \in \Xi$. Dempster's rule is extended to the general case simply by renormalizing the probabilities that are assigned to states in Ξ under the independence model, to allow for the fact that states outside Ξ are impossible. The multivalued mapping from Ξ to Ω then induces the probability assignment

$$m^{x,y}(C) = \rho^{-1} \sum_{A_i \cap B_j = C} m^x(A_i) m^y(B_j)$$

for non-empty sets C , where the normalizing constant ρ is the sum of all terms $m^x(A_i) m^y(B_j)$ for which $A_i \cap B_j$ is non-empty.

This is the general version of Dempster's rule.²⁰ It is justified provided the joint probabilities are precise and factorize as $P(\psi_i, \phi_j | x, y) = \rho^{-1} P(\psi_i | x) P(\phi_j | y)$ whenever $(\psi_i, \phi_j) \in \Xi$. Again, it does not appear that this condition will often be reasonable. To see that the answers produced by Dempster's rule may be unreasonable, it suffices to examine a special case, Dempster's rule of conditioning.

5.13.9 Dempster's rule of conditioning

Suppose that the second piece of evidence, y , is a reliable observation that event B has occurred, where B is a subset of Ω . Then $m^y(B) = 1$. When x and y are combined using Dempster's rule, we obtain

$$m^{x,y}(D) = \rho^{-1} \sum_{A_i \cap B = D} m^x(A_i), \text{ where } \rho = \bar{P}^x(B).$$

This special case of the rule of combination is known as **Dempster's rule of conditioning**.²¹ The combined probability assignment $m^{x,y}(D)$, which is usually written as a conditional probability $m^x(D|B)$, represents Your beliefs about Ω after learning (in addition to x) that B has occurred. The rule can be written most simply in terms of the corresponding upper probabilities, as $\bar{P}^x(D|B) = \bar{P}^x(D \cap B) / \bar{P}^x(B)$. (Compare with Bayes' rule.)

Dempster's rule of conditioning is reasonable provided You accept the probability judgements $P(\psi_i | x, y) = \rho^{-1} P(\psi_i | x)$, for all states ψ_i that are consistent with B . This means that learning that B has occurred does not change Your relative probabilities for those states that are consistent with B . Again, these judgements will be unreasonable in many problems, because the knowledge that B has occurred provides information about ψ through the multivalued mapping. (Loosely, it supports those states ψ_i for which A_i is contained in B , relative to those for which A_i lies mostly outside B .) In the following examples it is incoherent to make these judgements for all possible observations B , and the use of Dempster's rule produces a sure loss.²²

5.13.10 The 'three prisoners' problem²³

Three prisoners a, b, c are due to be executed. The authorities decide to reprieve one prisoner, who is chosen at random (each with chance $\frac{1}{3}$). In an attempt to gain information about his own prospects, prisoner a asks the governor to name one of the other two prisoners who has failed to win reprieve. The governor agrees to do so. (We assume that his answer is wholly reliable.) What will prisoner a learn about his prospects for survival?

Consider the possibility space $\Omega = \{ab, ac, bc, cb\}$, where st denotes the outcome that prisoner s is chosen to be reprieved and the governor states that t will be executed. Let $\Psi = \{\psi_a, \psi_b, \psi_c\}$, where ψ_s denotes the state in which s is reprieved. On the basis of his knowledge (x) about the random choice, prisoner a assigns precise probability $\frac{1}{3}$ to each state in Ψ . There is a multivalued mapping A^x from Ψ to Ω , defined by

$$A^x(\psi_a) = \{ab, ac\}, A^x(\psi_b) = \{bc\}, A^x(\psi_c) = \{cb\}.$$

This generates m^x which assigns probability $\frac{1}{3}$ to each of the sets $\{ab, ac\}$, $\{bc\}$ and $\{cb\}$.

The imprecision of this probability model simply reflects a 's ignorance about how the governor will answer his question when both b and c are to be executed: he does not know how to distribute the probability $\frac{1}{3}$ between the possibilities ab and ac . According to the standard analysis of this problem, as well as the analysis given in section 6.4.4, the imprecision of the model leads to imprecision in a 's posterior probabilities after he hears the governor's answer. We will show, however, that Dempster's rule produces precise posterior probabilities.

Let $B = \{ab, cb\}$ denote the event that the governor names b , and $C = \{ac, bc\}$ the event that he names c . Initially, a assigns precise probability $\frac{1}{3}$ to the event $R = \{ab, ac\}$, that he will be reprieved. Suppose that a uses Dempster's rule of conditioning to update his probability for R after learning from the governor that B (or C) has occurred. The probability assignments determined by Dempster's rule are $m^x(\{ab\}|B) = m^x(\{cb\}|B) = \frac{1}{2}$, and $m^x(\{ac\}|C) = m^x(\{bc\}|C) = \frac{1}{2}$. Hence $\bar{P}^x(R|B) = \bar{P}^x(R|C) = \bar{P}^x(R|B \cap C) = \frac{1}{2}$. Irrespective of whether the governor names b or c , a 's updated probability of reprieve will be precisely $\frac{1}{2}$.²⁴ The use of Dempster's rule apparently improves his prospects of survival!

The precise posterior assessments $P(R|B) = \frac{1}{2}$ and $P(R|C) = \frac{1}{2}$ are clearly inconsistent with the prior assessment $P(R) = \frac{1}{3}$. Assuming that You adopt these probabilities as betting rates, an observer could exploit Your rates to make a sure gain by initially betting that a will be reprieved and later betting that a will be executed. Thus Dempster's rule produces a 'sure loss'.

5.13.11 An indeterminate integer

Here is another, more extreme, example of a sure loss from Dempster's rule. Suppose that ω is an integer between 1 and 99. You know that the first digit of ω is chosen by a random mechanism which gives equal chances to the 10 digits 0, 1, ..., 9, but Your only information about the second digit is that it is non-zero. Hence Your probability assignment is $m(A_i) = 0.1$ for each of the sets $A_i = \{10i + 1, 10i + 2, \dots, 10i + 9\}$, where $i = 0, 1, \dots, 9$.

Now suppose that You are told that the first non-zero digit of ω is j , so You learn that ω belongs to the set $B_j = \{j\} \cup A_j$. Using Dempster's rule to condition on B_j , we obtain $m(\{j\}|B_j) = m(A_j|B_j) = 0.5$, for each $j = 1, 2, \dots, 9$. The set A_0 has precise probability 0.5 after updating by Dempster's rule, whatever the value of j , whereas initially A_0 had precise probability 0.1. Again, these assessments incur sure loss. Application of Dempster's rule is unreasonable because observation of B_j supports A_j (which is contained in B_j) more than A_0 (which intersects B_j at a single point).

5.13.12 Behavioural interpretation of belief functions

In the preceding examples, the updated probabilities defined by Dempster's rule are incoherent with the initial probabilities. Together they incur sure loss. Provided both the initial and updated probabilities have a behavioural interpretation, avoiding sure loss seems to be a necessary condition of rationality, and the probabilities produced by Dempster's rule seem to be wholly unreasonable. Do belief functions have such a behavioural interpretation?

Belief functions are a special type of lower probability, so they can be given the same behavioural interpretation. The conditional belief function $P(\cdot|B)$ is to be adopted as an updated (unconditional) belief function after B is observed, so it too can be given a behavioural interpretation. Moreover, it is clear from section 4.3.5 that any belief function obtained from a multivalued mapping inherits the behavioural meaning of the underlying precise probabilities. Both the initial and updated belief functions can be obtained from multivalued mappings.

It is possible, of course, to reject the behavioural interpretation and the model of multivalued mappings, but then the meaning of belief functions becomes quite unclear.²⁵ Anyway, Shafer (1982a) does accept the behavioural interpretation of belief functions: 'It seems to me that the degrees of belief constructed within the theory of belief functions and those constructed within the theory of lower probabilities can equally well be used as betting rates.'²⁶ That is true as long as attention is restricted to unconditional

probabilities, because belief functions are coherent lower probabilities, but not when conditional probabilities are constructed by Dempster's rule.

In fact, there is a unique rule of conditioning which generates conditional probabilities that are coherent with given initial probabilities. That is the generalized Bayes' rule (GBR), which is derived in the next chapter (section 6.4). The conditional probabilities defined by Dempster's rule are always at least as precise, and typically more precise, than those defined by the GBR.²⁷

5.13.13 Conclusion

The Dempster–Shafer theory of belief functions is, from our point of view, an important contribution to a theory of coherent upper and lower probabilities. We especially welcome the emphasis on constructing probabilities through careful modelling of evidence. The theory provides some useful methods for constructing belief functions, especially through multivalued mappings, and these can be used more generally as a source of coherent lower probabilities (see section 4.3.5). Greater generality is needed because the class of belief functions is not large enough to model all reasonable types of uncertainty.

Dempster's rule of combination has a central role in the theory, in combining evidence and updating probabilities, that is incompatible with an approach based on coherence. Dempster's rule can be justified in some problems through precise judgements of conditional independence, but these judgements seem to be unreasonable in most cases. In order to obtain probabilities from the combined evidence (x, y) through a multivalued mapping, it is necessary to make some judgements about the joint probabilities $P(\psi, \phi|x, y)$, either to justify the use of Dempster's rule or to obtain different answers. Without such judgements, Dempster's rule is unreliable.²⁸

In simple problems of conditioning, the use of Dempster's rule can produce a sure loss and is clearly unreasonable. In order to define coherent conditional probabilities You must use a different rule, the generalized Bayes' rule described in the following chapter.

CHAPTER 6

Conditional previsions

This chapter introduces the concept of conditional prevision, on which the theory of statistical reasoning in the following chapters is based. We suppose that conditional lower previsions, real numbers denoted by $\underline{P}(X|B)$, are assessed for various gambles X , and for various conditioning events B which belong to a partition \mathcal{B} of Ω . The notations $\underline{P}(\cdot|B)$, $\underline{P}(X|\mathcal{B})$ and $\underline{P}(\cdot|\mathcal{B})$ are used to represent collections of conditional previsions $\underline{P}(X|B)$.

As with unconditional previsions, the two main issues in this chapter are coherence (sections 6.1–6.6) and natural extension (sections 6.7–6.11). The coherence axioms, which relate conditional previsions $\underline{P}(\cdot|\mathcal{B})$ to unconditional previsions \underline{P} , are based on two fundamental principles, the updating principle (6.1.6) and the conglomerative principle (6.3.3). If \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent, we can try to extend them to larger domains in a coherent way. The main difference from the earlier theory of natural extension is that, when \mathcal{B} is infinite, coherent extension may not be possible.

In the special case where both \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are linear previsions, the theory in this chapter can be compared with the standard theory of conditional probability (due to Kolmogorov) and with de Finetti's theory. These approaches are compared in sections 6.5.8, 6.8 and 6.9. When the partition \mathcal{B} is finite and $\underline{P}(B) > 0$ for all B in \mathcal{B} , our coherence axioms reduce to Bayes' rule and the three approaches are essentially the same. In other cases, our coherence conditions are stronger than those of de Finetti and Kolmogorov. Like Kolmogorov, we require the conglomerative condition $\underline{P}(X) = \underline{P}(\underline{P}(X|\mathcal{B}))$, although we regard this as a consistency requirement whereas Kolmogorov adopts it as an implicit definition of conditional probability. De Finetti rejects this condition when \mathcal{B} is infinite. We also support Kolmogorov's assumption that \underline{P} is countably additive, which is rejected by de Finetti. In our theory, like de Finetti's but unlike Kolmogorov's, conditional probabilities are not defined in terms of unconditional ones, $\underline{P}(\cdot|B)$ can be defined when B has probability zero, $\underline{P}(\cdot|B)$ is required to be a coherent prevision, and all models are required to have coherent extensions to larger domains.

Here is an outline of the chapter. Two interpretations of conditional previsions are compared in section 6.1. They can be interpreted either as buying prices for gambles contingent on B , or as updated buying prices that You intend to adopt if You learn only that B has occurred. The updating principle implies that the two buying prices should be equal.

The basic notation for conditional previsions is introduced in section 6.2. Each $\underline{P}(\cdot|B)$ is required to be coherent as an unconditional lower prevision, with $\underline{P}(B|B) = 1$. In section 6.3 we justify further conditions which characterize coherence of conditional previsions $\underline{P}(\cdot|\mathcal{B})$ with unconditional previsions \underline{P} . One important consequence of coherence is the generalized Bayes rule (GBR), introduced in section 6.4, which implies that $\underline{P}(X|B)$ is determined by \underline{P} when $\underline{P}(B) > 0$, as the unique value of μ such that $\underline{P}(B(X - \mu)) = 0$. The GBR reduces to Bayes' rule when \underline{P} is linear.

In section 6.5, simple axioms are given to characterize coherence in various special cases. When \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are defined on the same domain, for example, coherence is equivalent to the GBR plus a conglomerative axiom. If also \mathcal{B} is finite, coherence is equivalent to the GBR alone. Some examples of conditional previsions are given in section 6.6, including linear–vacuous mixtures, constant odds-ratio models, zero–one valued probabilities, ‘uniform distributions’ on the positive integers and other non-conglomerable models.

In section 6.7 we turn to the problem of coherently extending \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ to larger domains, and examine the special case where \underline{P} is defined only on \mathcal{B} -measurable gambles. In this case, coherent extension is always possible.

The last four sections are concerned with the problem of defining conditional previsions $\underline{P}(\cdot|\mathcal{B})$ that are coherent with a lower prevision \underline{P} . That is possible if and only if \underline{P} satisfies a condition of \mathcal{B} -conglomerability, studied in section 6.8. We argue that the coherence axioms in Chapter 2 should be strengthened to include conglomerability. Section 6.9 investigates the connection between conglomerability and countable additivity. The results strengthen the arguments against the Bayesian requirement that previsions be linear, and against the sensitivity analysis interpretation of lower previsions. Linearity of \underline{P} is incompatible with the two requirements that it can be coherently extended to larger domains, and that it always has coherent conditional previsions.

The problem of conditioning on an event B that has probability zero is considered in section 6.10. Borel's paradox shows that conditional previsions are indeterminate unless further information is provided. One way to do this is to regard B as a limit of conditioning events with positive probability. This justifies a version of Bayes' rule for conditioning on a continuous variable, but only when substantive assumptions are made about the precision of measurement.

Finally, section 6.11 concerns the role of the GBR in updating beliefs after receiving new evidence. It is regarded as one amongst many possible updating strategies. Provided \underline{P} is assessed retrospectively after receiving the new evidence, the GBR can always be applied to impose constraints on the updated beliefs. But often these constraints are nearly vacuous, and then other updating strategies may be more useful than the GBR.

6.1 Updated and contingent previsions

The first issue is the interpretation of conditional previsions. In this section we describe two interpretations, discuss their uses in updating, assessment and statistical inference, and consider the relationship between them.

6.1.1 Two interpretations of conditional previsions

It is important to distinguish between two kinds of conditional previsions, which we will call updated previsions and contingent previsions.¹ Roughly, updated previsions describe Your present commitments to update Your beliefs if You happen to observe just the event B . Contingent previsions describe Your present dispositions to accept gambles which are zero outside B , or called off if B fails to occur. The two interpretations will be defined more carefully in the following paragraphs.²

First consider **contingent previsions**. Say that a gamble Y is **contingent** on an event B when Y is zero outside B , so $Y = BY$. The contingent interpretation of $\underline{P}(X|B)$ is that it is a supremum buying price for X contingent on B . In other words, You are currently willing to accept any contingent gamble of the form $B(X - \mu)$, provided μ is less than $\underline{P}(X|B)$. Thus a contingent prevision $\underline{P}(\cdot|B)$ models Your current attitudes to gambles contingent on B .³

The second interpretation of $\underline{P}(X|B)$, as an **updated prevision**, requires somewhat more care. We will apply it when the following conditions are satisfied. Suppose that an observation will be made between now and some later time, and that the possible observations can be identified with subsets of Ω which form a partition. Observation of a set B establishes just that the true state ω is in B . Then $\underline{P}(X|B)$ describes the attitudes to gambles $X - \mu$ that You are now committed to adopt after You observe B , provided You obtain no other relevant information about Ω .⁴ Specifically, You are prepared to pay up to $\underline{P}(X|B)$ for X after You make the observation, if You learn only that the true state is in B . Thus $\underline{P}(\cdot|B)$ models the updated beliefs You intend to adopt if You observe only B .

The assumption that the conditioning events B form a partition of Ω is quite restrictive, but it does hold in the important case of statistical observations, the subject of later chapters. The results of this chapter do not apply

6.1 UPDATED AND CONTINGENT PREVISIONS

directly to statistical problems, which usually involve several kinds of conditioning events, because here we restrict attention to conditioning events which form a single partition of Ω . This restriction seems necessary if all conditional previsions are to be interpreted as updated previsions. Updating is more difficult without the restriction because further probability assessments are needed.⁵

When all conditional previsions are interpreted as contingent previsions, the restriction to partitions is unnecessary and any collection of conditioning events can be considered, using the more general formulation in section 7.1. The ‘contingent’ interpretation is evidently wider than the ‘updating’ interpretation in that it is not restricted to problems where the conditioning events can be regarded as possible outcomes of an experiment.

Both contingent and updated previsions model Your current behavioural dispositions. The difference is that the dispositions described by contingent previsions can be displayed in Your current behaviour, whereas those described by updated previsions can be displayed only at a later time, under the special conditions that B is observed and You learn nothing more about Ω .⁶

6.1.2 Updating beliefs

The most obvious use of updated previsions is in actually revising (updating) beliefs when new evidence is obtained. The updated previsions $\underline{P}(X|B)$ describe Your present commitments to revise Your beliefs after obtaining the new evidence that B has occurred. These should be distinguished from the unconditional lower previsions $\underline{P}^B(X)$ that You adopt after observing B . We will assume that Your commitments are reliable, in the sense that You do become willing to pay up to the updated prevision $\underline{P}(X|B)$ for X if You observe B . Then $\underline{P}^B(X) \geq \underline{P}(X|B)$. But often Your new previsions \underline{P}^B will be more precise than $\underline{P}(\cdot|B)$, so $\underline{P}^B(X) > \underline{P}(X|B)$ for some X , because observation of B leads You to think more carefully about Your attitudes to gambles contingent on B and to make further assessments.⁷

When Your current assessments of the contingent gambles are sufficiently thorough, we might expect \underline{P}^B to agree with $\underline{P}(\cdot|B)$. More generally, $\underline{P}(\cdot|B)$ will still be a correct description of Your new beliefs after observing B , but it will be incomplete in the sense of section 2.10.3. It is reasonable to expect the updated previsions $\underline{P}(\cdot|B)$ to be coherent with Your unconditional previsions \underline{P} , since both represent current dispositions, but less reasonable to expect \underline{P} to be coherent with \underline{P}^B .

Several rules have been suggested for updating beliefs in the light of new evidence. A strict Bayesian would use Bayes’ rule to condition on the new evidence. Those who favour other theories of conditioning might be

committed to other rules, such as Dempster's rule (section 5.13.9) or Jeffrey's rule (section 6.11.8). Anyone who intends to use a particular updating rule has, according to our interpretation, the updated previsions determined by the rule, even though he may not have computed these explicitly.

We are interested in which of these updating rules is coherent. On our interpretation of updated previsions, that question reduces to asking when Your current dispositions (which include dispositions to update beliefs) are coherent. We will see that, under some restrictive assumptions, one updating rule is uniquely coherent. This is the generalized Bayes rule. Coherent updated previsions $\underline{P}(X|B)$ are uniquely determined, through this rule, by the unconditional previsions \underline{P} .⁸

6.1.3 Assessment

One difficulty in using the generalized Bayes rule is that the updated previsions it defines may be highly imprecise. That happens when Your unconditional previsions for gambles contingent on the observed event B are imprecise, which may be unavoidable when there are too many possible observations to consider in advance. After observing B , You may want to construct a new model \underline{P}^B that is more precise than $\underline{P}(·|B)$, by making further assessments.

This suggests that the generalized Bayes rule should be regarded as just one of many possible assessment strategies for updating beliefs. It is the only reasonable strategy which uses just the unconditional assessments \underline{P} , but other strategies, which use other assessments, are equally valid and may lead to greater precision.

More generally, conditional previsions have an important role in the assessment process, of which updating beliefs by conditioning is only a part. Any assessments of conditional and unconditional previsions have implications, through coherence and natural extension, for other previsions. As a simple example, assessments of precise probabilities $P(A|B)$ and $P(B)$ determine the precise probability $P(A \cap B) = P(A|B)P(B)$ by natural extension. Even when You are interested only in unconditional probabilities such as $P(A \cap B)$, it may be useful to assess conditional probabilities such as $P(A|B)$. These might be interpreted as either contingent or updated probabilities, but the contingent interpretation will usually be simpler because it does not require that B can be embedded in a partition of 'possible observations'.

6.1.4 Statistical inference

The role of conditional previsions in both updating and assessment can be illustrated by problems of statistical inference. Suppose that the observed

6.1 UPDATED AND CONTINGENT PREVISIONS

outcome x of an experiment is generated by an unknown sampling distribution $P(·|\theta)$, where the possible observations x belong to a sample space \mathcal{X} , and the possible parameter values θ belong to a parameter space Θ . Prior to the experiment, Your beliefs about the parameter are modelled by a lower prevision \underline{P} . The prior \underline{P} and sampling models $P(·|\theta)$ then generate prior probabilities concerning events in the product space $\Theta \times \mathcal{X}$, by natural extension. This is one way of assessing predictive probabilities concerning the outcome of the experiment.⁹

Another important problem is to update beliefs about the parameter after observing the outcome x . That can also be done by natural extension of the prior and sampling models, using the generalized Bayes rule to compute the updated previsions $\underline{P}(·|x)$.

In this case there are two kinds of conditional previsions, the sampling models $P(·|\theta)$ conditional on the unknown parameter θ , and the posterior previsions $\underline{P}(·|x)$ conditional on the observation x . It is natural to interpret $\underline{P}(·|x)$ as an updated prevision, since x is actually observed and beliefs need to be updated after the observation. Since θ is usually unobservable, it is more natural to interpret $P(·|\theta)$ as a contingent prevision.¹⁰

6.1.5 Equality of updated and contingent previsions

Despite the distinction we have made between updated and contingent previsions, it is clear that the two concepts are closely related. We will now argue that, in order to be consistent, updated and contingent previsions should agree. The reason is that buying a gamble X for μ contingent on B has exactly the same effect as buying X for μ after B occurs.

The argument is simpler in terms of desirability. Call a gamble Z **B -desirable** if You intend to accept Z provided You observe just the event B . Then B -desirability is related to updated previsions in the same way that ordinary desirability is related to unconditional previsions, and that the desirability of contingent gambles is related to contingent previsions. In fact, the agreement of contingent and updated previsions is implied by (and roughly equivalent to) the following general principle, which states that B -desirability is equivalent to desirability of gambles contingent on B .

6.1.6 Updating principle¹¹

Any gamble Z is B -desirable if and only if BZ is desirable.

This is a condition of consistency amongst Your current dispositions. The principle is compelling because, assuming that we will observe one of the sets in a partition \mathcal{B} that includes B , I can produce exactly the same outcome by giving You Z after the observation, if B is observed, as by

giving You BZ now. In both cases You receive the gamble BZ .¹² The two dispositions mentioned in the updating principle have the same effect.¹³

To see that the updating principle implies equality of updated and contingent previsions, let $Z = X - \mu$ be a gamble in which You buy X for price μ . Your updated prevision $\underline{P}_u(X|B)$, which is the supremum price μ for which $X - \mu$ is B -desirable, agrees with Your contingent prevision $\underline{P}_c(X|B)$, which is the supremum price μ for which $B(X - \mu)$ is desirable.

A strong kind of inconsistency occurs when the intervals $[\underline{P}_u(X|B), \bar{P}_u(X|B)]$ and $[\underline{P}_c(X|B), \bar{P}_c(X|B)]$ are disjoint, where $\bar{P}(X|B) = -\underline{P}(-X|B)$ is the conjugate upper prevision. To see that, suppose $\underline{P}_c(X|B) > \lambda > \mu > \bar{P}_u(X|B)$. By interpretation of the contingent and updated previsions, You are currently willing to accept the contingent gamble $B(X - \lambda)$, and to accept $\mu - X$ later if You observe B . The two gambles together produce a net loss of $\lambda - \mu$ if B occurs, and zero otherwise.¹⁴ Unless both gambles are called off, You must lose. This happens also when $\underline{P}_u(X|B) > \bar{P}_c(X|B)$. A Bayesian, whose updated and contingent previsions are precise, can be made to suffer this kind of loss unless the two previsions are equal.¹⁵

6.1.7 Football example

To illustrate the updating principle, consider a football game with $\Omega = \{W, D, L\}$. Let $B = \{W, L\}$ be the event that the game is not drawn. Before learning the full result of the game, You can find out whether or not B occurred by listening to a radio report which lists all the drawn games.¹⁶ Consider a judgement that W is more probable than L , meaning that $Z = W - L$ is desirable. The updating principle implies that You should make this judgement before hearing the radio report if and only if You intend to make the same judgement after hearing that the game was not drawn. That seems reasonable because the new information just eliminates the possibility D , on which the gamble Z is zero. It seems inconsistent to allow Your comparison between W and L to depend on whether D has been eliminated.

It is important, for the updating principle to be valid, that You learn *only* that B has occurred. If You learned that the game was a type of cup-tie for which D is not possible, then that provides extra information which might alter Your judgements. (W may be more probable than L in cup-ties, but not in other games.)

Another reason for failure of the updating principle is that You may be reluctant to make judgements of B -desirability, because they commit You to act in a certain way at a later time, after You make the observation. You may intend to re-assess Your evidence before the later time, and perhaps revise some of Your current judgements.¹⁷ So You might judge that $W - L$

6.2 SEPARATE COHERENCE

is desirable now, but not that it is B -desirable, because You intend to study the game more carefully before You hear the radio report.

It is important that probability assessments should be open to reappraisal and revision from time to time. This means that updated previsions may be less precise than contingent previsions, and the updating principle may be sometimes too stringent. But the principle is reasonable in those cases where Your probability model is based on a reliable analysis of the current evidence. If Your judgement that W is more probable than L is unreliable, and might be revised on further analysis of the same evidence, You should not make it. As long as You restrict Your judgements to those that are clearly justified, it is reasonable to expect You to maintain these at later times.

6.1.8 Conclusion

The updating principle is essential for relating initial beliefs to updated beliefs. We will see that it implies (and is roughly equivalent to) an updating rule, the generalized Bayes rule, for defining updated previsions in terms of initial, unconditional previsions. The principle seems justified when Your initial assessments are reliable, and it then enables You to automatically update Your beliefs in the light of the new observation B , without having to re-assess Your evidence. Henceforth we will assume that the updating principle holds, so that contingent and updated previsions are equal, but we will identify exactly where it is used in developing a theory of coherence.

6.2 Separate coherence

In this section we characterize coherence of conditional previsions in the case where the conditioning events form a partition of Ω . First we introduce the basic notation for conditional previsions to be used in this and the following chapters.

6.2.1 Notation

Let \mathcal{B} denote a **partition** of Ω , so \mathcal{B} is a class of non-empty, pairwise-disjoint subsets whose union is Ω .¹ Suppose that real numbers $\underline{P}(X|B)$ are specified for B in \mathcal{B} and all gambles X in some domain $\mathcal{H}(B)$, which includes the gamble B . Then $\underline{P}(\cdot|B)$ can be regarded as a lower prevision with domain $\mathcal{H}(B)$. Let $G(X|B) = B(X - \underline{P}(X|B))$ denote the **marginal gamble** on X contingent on B . This is the gamble in which You pay price $\underline{P}(X|B)$ for X , but the transaction is called off unless B occurs.

We can now define coherence of a collection of conditional previsions $\underline{P}(X|B)$, specified for various gambles X and conditioning events B .

6.2.2 Definition

Conditional previsions $\underline{P}(X|B)$, defined for B in \mathcal{B} and X in $\mathcal{H}(B)$, are **separately coherent** when, for every conditioning event B , $\underline{P}(\cdot|B)$ is a coherent lower prevision on domain $\mathcal{H}(B)$ and $\underline{P}(B|B) = 1$.²

6.2.3 Justification

This requirement of separate coherence can be justified by applying the earlier argument for coherence (in section 2.5.2) to the marginal gambles $G(X|B)$. Suppose that $\underline{P}(\cdot|B)$ is not coherent. Then there are gambles X_0, X_1, \dots, X_n in $\mathcal{H}(B)$, $m \geq 0$ and $\delta > 0$ such that $Z = \sum_{j=1}^n (G(X_j|B) + \delta B) \leq mG(X_0|B) - \delta B$. Now Z is effectively desirable, either now (under the contingent interpretation) or after the observation is made (under the updating interpretation), because it is a sum of desirable gambles in which You pay less than $\underline{P}(X_j|B)$ for X_j contingent on B . If $m = 0$ then Z produces a sure loss when B occurs, and is otherwise zero. If $m > 0$ then $mG(X_0|B) - \delta B$ is desirable, which means that You are effectively willing to pay more than $\underline{P}(X_0|B)$ for X_0 contingent on B (or after observing B). Thus the conditional previsions are inconsistent.

It is therefore reasonable to require that each $\underline{P}(\cdot|B)$ be a coherent lower prevision. This implies that $\underline{P}(B|B) \leq 1$. The extra requirement $\underline{P}(B|B) \geq 1$ is clearly justified, because when $\mu < 1$ the contingent gambles $B(B - \mu) = B(1 - \mu)$ are non-negative and desirable.

The condition 6.2.2 is termed *separate* coherence because it simply requires that the separate lower previsions $\underline{P}(\cdot|B)$ are each coherent. There is nothing in the definition to connect the lower previsions for different conditioning events. Separate coherence seems to be sufficiently strong to characterize consistency of conditional previsions, provided the conditioning events belong to a single partition (so that they are indeed ‘separate’). It needs to be strengthened when the conditioning events may overlap (see section 7.1), or when unconditional previsions are assessed as well as conditional ones (see section 6.3). Before studying the latter problem, it is useful to examine some of the consequences of separate coherence.

First we show that there is no loss of generality in assuming that the domains $\mathcal{H}(B)$ are the same for all conditioning events B . That is a consequence of the next lemma, which shows that $\underline{P}(X|B)$ can depend only on the restriction of X to the set B .

6.2.4 Lemma

Suppose $\underline{P}(\cdot|B)$ is a coherent lower prevision on domain $\mathcal{H}(B)$ and $\underline{P}(B|B) = 1$. If X and Y are in $\mathcal{H}(B)$ and $BX = BY$ then $\underline{P}(X|B) = \underline{P}(Y|B)$.

6.2 SEPARATE COHERENCE

Proof. By coherence, $\sup[n(B - 1) + Y - \underline{P}(Y|B) - X + \underline{P}(X|B)] \geq 0$ for any $n > 0$, using $\underline{P}(B|B) = 1$. By choosing n sufficiently large, $\sup_{\omega \in B} [Y - \underline{P}(Y|B) - X + \underline{P}(X|B)] \geq 0$. Hence $\underline{P}(X|B) \geq \underline{P}(Y|B)$, using $BX = BY$. The result follows by interchanging X and Y . ♦

Separately coherent previsions $\underline{P}(\cdot|B)$ can now be extended to a common domain \mathcal{H} , as follows. First enlarge each $\mathcal{H}(B)$ to include all constant gambles μ , with $\underline{P}(\mu|B) = \mu$ to preserve separate coherence. Then define \mathcal{H} to consist of all gambles Y that (for every B in \mathcal{B}) agree on B with some gamble in $\mathcal{H}(B)$. If Y agrees with X_B on B , then $\underline{P}(Y|B) = \underline{P}(X_B|B)$ is well-defined by the lemma. Thus \mathcal{H} is constructed by ‘piecing together’ the gambles BX_B , where $X_B \in \mathcal{H}(B)$, in all possible ways.³

Henceforth we will assume that the conditional previsions $\underline{P}(\cdot|B)$ are separately coherent and defined on a common domain \mathcal{H} . This assumption greatly simplifies the notation, as follows.

6.2.5 Notation

A gamble X is called \mathcal{B} -measurable when X is constant on each set B in \mathcal{B} . Let $\mathcal{G}(\mathcal{B})$ denote the class of \mathcal{B} -measurable gambles. For X in \mathcal{H} , let $\underline{P}(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} B\underline{P}(X|B)$ denote the \mathcal{B} -measurable gamble with reward $\underline{P}(X|\mathcal{B})(\omega) = \underline{P}(X|B)$ when $\omega \in B$. Let $\bar{P}(X|\mathcal{B}) = -\underline{P}(-X|\mathcal{B}) = \sum_{B \in \mathcal{B}} B\bar{P}(X|B)$, and $G(X|\mathcal{B}) = X - \underline{P}(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} G(X|B)$.

The gamble $G(X|\mathcal{B})$, in which You pay the uncertain price $\underline{P}(X|\mathcal{B})$ for X , has an important role in the following theory.⁴ It can be regarded as a **two-stage gamble** in which (i) You observe B and pay price $\underline{P}(X|B)$, (ii) You observe ω (in B) and receive $X(\omega)$.

We will use $\underline{P}(\cdot|\mathcal{B})$ to denote the function from \mathcal{H} into $\mathcal{G}(\mathcal{B})$ which summarizes the collection of coherent lower previsions $\{\underline{P}(\cdot|B) : B \in \mathcal{B}\}$. We will assume that $\underline{P}(\cdot|\mathcal{B})$ is separately coherent, meaning that all the $\underline{P}(\cdot|B)$ are separately coherent.

Using this notation, the properties of separately coherent conditional previsions can be summarized as follows. (These are simple consequences of the coherence properties 2.6.1 and Definition 6.2.2.) Properties (j) and (l) are especially useful.

6.2.6 Consequences of separate coherence

Suppose that $\underline{P}(\cdot|\mathcal{B})$ is separately coherent. Let $\bar{P}(\cdot|\mathcal{B})$ be the conjugate upper conditional prevision, and let $\mathcal{G}^+(\mathcal{B})$ denote the class of all non-negative \mathcal{B} -measurable gambles. The following properties hold whenever the

conditional previsions are defined:

- (a) $\sup\{X(\omega): \omega \in B\} \geq \bar{P}(X|B) \geq \underline{P}(X|B) \geq \inf\{X(\omega): \omega \in B\}$ for all $B \in \mathcal{B}$,
and $\sup X \geq \bar{P}(X|\mathcal{B}) \geq \underline{P}(X|\mathcal{B}) \geq \inf X$
- (b) when $\lambda > 0$, $\underline{P}(\lambda X|\mathcal{B}) = \lambda \underline{P}(X|\mathcal{B})$, $\bar{P}(\lambda X|\mathcal{B}) = \lambda \bar{P}(X|\mathcal{B})$
- (c) $\underline{P}(X + Y|\mathcal{B}) \geq \underline{P}(X|\mathcal{B}) + \underline{P}(Y|\mathcal{B})$, $\bar{P}(X + Y|\mathcal{B}) \leq \bar{P}(X|\mathcal{B}) + \bar{P}(Y|\mathcal{B})$
- (d) when $\mu \in \mathbb{R}$, $\underline{P}(\mu|\mathcal{B}) = \bar{P}(\mu|\mathcal{B}) = \mu$
- (e) if $X(\omega) = \mu$ for all $\omega \in B$ then $\underline{P}(X|B) = \bar{P}(X|B) = \mu$
- (f) when $Y \in \mathcal{G}(\mathcal{B})$, $\underline{P}(Y|\mathcal{B}) = \bar{P}(Y|\mathcal{B}) = Y$
- (g) if $X \geq Y$ then $\underline{P}(X|\mathcal{B}) \geq \underline{P}(Y|\mathcal{B})$ and $\bar{P}(X|\mathcal{B}) \geq \bar{P}(Y|\mathcal{B})$
- (h) if $X \geq Y$ and $Y \in \mathcal{G}(\mathcal{B})$ then $\underline{P}(X|\mathcal{B}) \geq Y$
- (i) if $Y \in \mathcal{G}(\mathcal{B})$ then $\underline{P}(X + Y|\mathcal{B}) = \underline{P}(X|\mathcal{B}) + Y$ and $\bar{P}(X + Y|\mathcal{B}) = \bar{P}(X|\mathcal{B}) + Y$
- (j) $\underline{P}(G(X|\mathcal{B})|\mathcal{B}) = 0$, $\bar{P}(X - \bar{P}(X|\mathcal{B})|\mathcal{B}) = 0$
- (k) if $X \geq 0$ then $\underline{P}(X|\mathcal{B}) \in \mathcal{G}^+(\mathcal{B})$
- (l) when $Y \in \mathcal{G}^+(\mathcal{B})$, $\underline{P}(YX|\mathcal{B}) = Y\underline{P}(X|\mathcal{B})$ and $\bar{P}(YX|\mathcal{B}) = Y\bar{P}(X|\mathcal{B})$
- (m) when $Y \in \mathcal{G}^+(\mathcal{B})$, $\underline{P}(YG(X|\mathcal{B})|\mathcal{B}) = 0$ and $\bar{P}(Y(X - \bar{P}(X|\mathcal{B}))|\mathcal{B}) = 0$
- (n) when $Y \in \mathcal{G}(\mathcal{B})$, $\underline{P}(YX|\mathcal{B}) \leq Y\underline{P}(X|\mathcal{B})$, $\bar{P}(YX|\mathcal{B}) \leq Y\bar{P}(X|\mathcal{B})$,
 $\bar{P}(YX|\mathcal{B}) \geq Y\bar{P}(X|\mathcal{B})$, $\bar{P}(YX|\mathcal{B}) \geq Y\underline{P}(X|\mathcal{B})$
- (o) when $Y \in \mathcal{G}(\mathcal{B})$, $\underline{P}(YG(X|\mathcal{B})|\mathcal{B}) \leq 0$ and $\bar{P}(Y(X - \bar{P}(X|\mathcal{B}))|\mathcal{B}) \geq 0$.

To simplify the ensuing theory, we will assume that the domain \mathcal{H} is a linear space.⁵ In that case, coherence is characterized by the axioms P1–P3 in section 2.3.3, and separate coherence is characterized by the following related axioms.

6.2.7 Theorem

Suppose $\underline{P}(\cdot|\mathcal{B})$ is defined on domain \mathcal{H} which is a linear space containing all gambles $B \in \mathcal{B}$. Then $\underline{P}(\cdot|\mathcal{B})$ is separately coherent if and only if it satisfies the three axioms, for all X and Y in \mathcal{H} ,

- (C1) $\underline{P}(X|B) \geq \inf\{X(\omega): \omega \in B\}$ when $B \in \mathcal{B}$ ⁶
- (C2) $\underline{P}(\lambda X|\mathcal{B}) = \lambda \underline{P}(X|\mathcal{B})$ when $\lambda > 0$
- (C3) $\underline{P}(X + Y|\mathcal{B}) \geq \underline{P}(X|\mathcal{B}) + \underline{P}(Y|\mathcal{B})$.

Proof. Separate coherence implies C1, by taking $Y = 0$ in the proof of Lemma 6.2.4, and also C2 and C3, using coherence of each $\underline{P}(\cdot|B)$. Conversely, the three axioms imply that $\underline{P}(\cdot|B)$ satisfies P1–P3 of section 2.3.3 and is therefore coherent. Hence $\underline{P}(B|B) \leq \sup B = 1$, and $\underline{P}(B|B) \geq 1$ by C1. ◆

6.3 COHERENCE WITH UNCONDITIONAL PREVISIONS

The **vacuous** conditional previsions $\underline{P}(X|B) = \inf\{X(\omega): \omega \in B\}$ satisfy these three axioms, and it is clear from C1 that these are the minimal separately coherent conditional previsions.

6.3 Coherence with unconditional previsions

In the previous section we characterized coherence of conditional previsions $\underline{P}(\cdot|\mathcal{B})$ defined for a partition \mathcal{B} . Next we introduce an unconditional lower prevision \underline{P} and examine when that is coherent with $\underline{P}(\cdot|\mathcal{B})$. This case, in which both \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are specified, is the most general case considered in the rest of this chapter. The following simplifying assumptions will be made throughout the chapter.

6.3.1 Assumptions

We assume that \underline{P} is an unconditional lower prevision defined on domain \mathcal{K} , and $\underline{P}(\cdot|\mathcal{B})$ is a conditional lower prevision defined on domain \mathcal{H} , with the following properties:

- (a) \mathcal{H} and \mathcal{K} are linear subspaces of $\mathcal{L}(\Omega)$ that contain all constant gambles.
- (b) If $Y \in \mathcal{H}$ then $BY \in \mathcal{H}$ for all $B \in \mathcal{B}$.¹
- (c) If $Y \in \mathcal{H}$ then $\underline{P}(Y|\mathcal{B}) \in \mathcal{H}$.
- (d) $\underline{P}(\cdot|\mathcal{B})$ is separately coherent on \mathcal{H} (i.e., satisfies the axioms in Theorem 6.2.7).
- (e) \underline{P} is coherent on \mathcal{K} (i.e., satisfies the axioms in section 2.3.3).

These assumptions are made to simplify the definition of coherence and the ensuing mathematical theory, and they are not as restrictive as they may appear. It was shown in section 6.2.4 that specified conditional previsions can be extended to a domain \mathcal{H} that satisfies properties (b) and (c), and contains all constant gambles, so there is no loss of generality in these assumptions. Coherence of \underline{P} and separate coherence of $\underline{P}(\cdot|\mathcal{B})$ are obviously necessary for coherence of the pair $\underline{P}, \underline{P}(\cdot|\mathcal{B})$. Assumptions (d) and (e) allow us to concentrate on the linkage between \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ that is needed for coherence. The only restrictive assumption is that the domains \mathcal{H} and \mathcal{K} are linear spaces. This assumption can be removed if the gambles $G(X)$ and $G(Y|\mathcal{B})$ in the following definitions are replaced by arbitrary finite sums $\sum_{j=1}^n G(X_j)$ and $\sum_{i=1}^k G(Y_i|\mathcal{B})$.²

6.3.2 Definition

Suppose that assumptions 6.3.1 hold. Recall the notation $G(X) = X - \underline{P}(X)$, $G(Y|\mathcal{B}) = Y - \underline{P}(Y|\mathcal{B})$, $G(W|B) = B(W - \underline{P}(W|B))$. Say that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$

avoid sure loss if

$$\sup[G(X) + G(Y|\mathcal{B})] \geq 0 \quad \text{whenever } X \in \mathcal{K} \quad \text{and} \quad Y \in \mathcal{H}.$$

Say that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent if

- (a) $\sup[G(X) + G(Y|\mathcal{B}) - G(Z)] \geq 0 \quad \text{and}$
- (b) $\sup[G(X) + G(Y|\mathcal{B}) - G(W|B)] \geq 0,$

whenever $X \in \mathcal{K}, Y \in \mathcal{H}, Z \in \mathcal{K}, W \in \mathcal{H}$ and $B \in \mathcal{B}$.

These conditions are similar to those studied in Chapter 2. Coherence is a stronger condition than avoiding sure loss in general, but the two conditions are equivalent when both \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are linear previsions. The two conditions hold automatically (by coherence of \underline{P} and separate coherence of $\underline{P}(\cdot|\mathcal{B})$) when either domain \mathcal{K} or \mathcal{H} is trivial, i.e., when \mathcal{K} contains only constant gambles or \mathcal{H} contains only \mathcal{B} -measurable gambles. This illustrates that the two conditions in the above definition are essentially concerned with the linkage between \underline{P} and $\underline{P}(\cdot|\mathcal{B})$.

Justification of these conditions relies on the updating principle (6.1.6) and the following important principle.

6.3.3 Conglomerative principle

If a gamble Z is B -desirable for every set B in the partition \mathcal{B} , then Z is desirable.

The conglomerative principle, like the updating principle, requires a type of consistency between current beliefs and current dispositions to update beliefs. It can be justified as follows. Suppose that Z is B -desirable for every B in \mathcal{B} . This means that You will be willing to accept Z after You observe some set in \mathcal{B} , whatever set You observe. Knowing this, You should be willing to accept Z now. (Again, the time at which gambles are accepted does not affect their value.)³

Failure of the conglomerative principle is especially serious when Z is B -desirable for every B in \mathcal{B} , but Z is strictly undesirable (i.e., $-Z - \delta$ is desirable for some positive δ). Then You are willing to accept $-Z - \delta$ now, and to accept Z after the experiment (whatever B is observed), from which You incur sure loss. This often happens when Bayes' rule (which is a consequence of the updating principle) is used to update a linear prevision that is not countably additive on the partition \mathcal{B} .⁴

A most important corollary of the conglomerative principle is that the two-stage gamble $G(Y|\mathcal{B}) = Y - \underline{P}(Y|\mathcal{B})$, where $\underline{P}(Y|\mathcal{B})$ is interpreted as an updated prevision, is almost desirable.⁵ To prove that, it suffices to prove $Z = G(Y|\mathcal{B}) + \delta$ is desirable when δ is positive. After observing B , Z is the gamble in which You pay $\underline{P}(Y|B) - \delta$ for Y .⁶ By interpretation of updated

6.3 COHERENCE WITH UNCONDITIONAL PREVISIONS

previsions (6.1.1), Z is B -desirable for every B in \mathcal{B} . So Z is desirable by the conglomerative principle.

The conglomerative principle and updating principle (6.1.6) together imply the following **contingent version of the conglomerative principle**: if BZ is desirable for every B in the partition \mathcal{B} , then Z is desirable.⁷ This follows from axiom D3 of section 2.2.3 when the partition is finite, since then $Z = \sum_{B \in \mathcal{B}} BZ$ is a finite sum of desirable gambles, but it has been disputed by some authors (notably de Finetti) in the case of an infinite partition.⁸ However, the updating and conglomerative principles seem to be just as compelling for infinite partitions as for finite ones. Indeed, the cardinality of \mathcal{B} seems to be irrelevant in the arguments given to support these principles.

It follows from the contingent version of the conglomerative principle that the gamble $G(Y|\mathcal{B})$ is almost desirable when $\underline{P}(Y|\mathcal{B})$ is interpreted as a contingent prevision. To see that, let $Z = G(Y|\mathcal{B}) + \delta$, and note that $BZ = B(Y - \underline{P}(Y|B) + \delta)$ is desirable by interpretation of the contingent prevision $\underline{P}(Y|B)$.

6.3.4 Justification of the coherence conditions

It is now straightforward to justify the coherence conditions in Definition 6.3.2, using the conglomerative principle (6.3.3) and the updating principle (6.1.6). The marginal gamble $G(X)$ is almost desirable, by interpretation of $\underline{P}(X)$, and $G(Y|\mathcal{B})$ is almost desirable, by the argument in section 6.3.3. Hence the sum $G(X) + G(Y|\mathcal{B})$ is almost desirable, by axiom D3 of section 2.2.3.

If the avoiding sure loss condition fails then there is a strictly desirable gamble $G(X) + G(Y|\mathcal{B}) + \delta$ which is uniformly negative. You are willing to pay a little less than $\underline{P}(X)$ for X , and a little less than $\underline{P}(Y|B)$ for Y after observing B , but You will certainly lose from these transactions. The dispositions represented by \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are irrational because they produce a sure loss.

If the coherence condition fails then there is a gamble of the form (a) $G(Z) - \delta$, or (b) $G(W|B) - \delta$, that is uniformly better than the strictly desirable gamble $G(X) + G(Y|\mathcal{B}) + \delta$. In case (a), You should be willing to pay price $\underline{P}(Z) + \delta$ for Z . Your unconditional prevision $\underline{P}(Z)$ is inconsistent with the other previsions, and can be increased by at least δ . In case (b), You should be willing to pay $\underline{P}(W|B) + \delta$ for W contingent on B . Your contingent prevision $\underline{P}(W|B)$ can be increased by δ , and the same is true of Your updated prevision (by the updating principle). Incoherence occurs just when one of Your conditional or unconditional lower previsions can be increased, using the information provided by the other previsions.

Because the updating principle implies that contingent and updated

previsions are equal, the coherence conditions apply under both interpretations of conditional previsions. However, the principles needed to justify coherence are quite different under the two interpretations. For contingent previsions, both avoiding sure loss and coherence rely on the contingent version of the conglomerative principle. This follows from axioms 2.2.3 if \mathcal{B} is finite, and then coherence is no stronger than the axioms studied in Chapter 2.

For updated previsions, only the conglomerative principle is needed to justify avoiding sure loss and part (a) of the coherence condition.⁹ The updating principle is needed to justify (b) of coherence in general, but not when the conditional revision $P(\cdot|\mathcal{B})$ is linear, nor when $\underline{P}(B) = 0$ for all B in \mathcal{B} .¹⁰ When \mathcal{B} is finite, coherence can be justified for updated previsions through the updating principle alone, without the conglomerative principle.

6.3.5 Consequences of coherence

Suppose that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ satisfy assumptions 6.3.1 and are coherent. The following properties hold whenever all previsions involved are defined. (Properties 1–3 are consequences of avoiding sure loss, and 4–8 follow from (a) of the coherence condition.)

1. $\underline{P}(X) \leq \sup \bar{P}(X|\mathcal{B})$, $\bar{P}(X) \geq \inf \underline{P}(X|\mathcal{B})$
2. $\underline{P}(X) \leq \bar{P}(\bar{P}(X|\mathcal{B}))$, $\bar{P}(X) \geq \underline{P}(\underline{P}(X|\mathcal{B}))$
3. $\underline{P}(X - \bar{P}(X|\mathcal{B})) \leq 0$, $\bar{P}(X - \underline{P}(X|\mathcal{B})) \geq 0$
4. $\underline{P}(X) \geq \inf \underline{P}(X|\mathcal{B})$, $\bar{P}(X) \leq \sup \bar{P}(X|\mathcal{B})$
5. $\underline{P}(X) \geq \underline{P}(\bar{P}(X|\mathcal{B}))$, $\bar{P}(X) \leq \bar{P}(\underline{P}(X|\mathcal{B}))$
6. $\underline{P}(X) \leq \bar{P}(\bar{P}(X|\mathcal{B}))$, $\bar{P}(X) \geq \bar{P}(\underline{P}(X|\mathcal{B}))$
7. $\underline{P}(X - \bar{P}(X|\mathcal{B})) \geq 0$, $\bar{P}(X - \underline{P}(X|\mathcal{B})) \leq 0$
8. If $Y \in \mathcal{G}(\mathcal{B})$ and $Y \geq 0$ then $\underline{P}(Y(X - \underline{P}(X|\mathcal{B}))) \geq 0$ and
 $\bar{P}(Y(X - \bar{P}(X|\mathcal{B}))) \leq 0$.
9. $\underline{P}(X) \leq \sup [B\underline{P}(X|B) + B^c\bar{P}(X|\mathcal{B})]$ and $\bar{P}(X) \geq \inf [B\bar{P}(X|B) + B^c\underline{P}(X|\mathcal{B})]$
10. $\underline{P}(B(X - \underline{P}(X|\mathcal{B}))) = 0$, $\bar{P}(B(X - \bar{P}(X|\mathcal{B}))) = 0$.
11. For all real μ , $\underline{P}(B(X - \mu))$ is non-negative (or non-positive) whenever $\underline{P}(X|B) - \mu$ is, and $\bar{P}(B)|\underline{P}(X|B) - \mu| \geq |\underline{P}(B(X - \mu))| \geq \underline{P}(B)|\underline{P}(X|B) - \mu|$. If $\underline{P}(B) > 0$, $\underline{P}(B(X - \mu))$ is positive (or negative or zero) whenever $\underline{P}(X|B) - \mu$ is.
12. If $\underline{P}(X|B) = 0$ then $\underline{P}(BX) = 0$; if $\bar{P}(X|B) = 0$ then $\bar{P}(BX) = 0$.
13. If $\underline{P}(X|B) \neq 0$ then $\underline{P}(B) \leq \underline{P}(BX)/\underline{P}(X|B) \leq \bar{P}(B)$.
If $\bar{P}(X|B) \neq 0$ then $\bar{P}(B) \leq \bar{P}(BX)/\bar{P}(X|B) \leq \bar{P}(B)$.
14. If $\underline{P}(X|B) \geq 0$ then $\underline{P}(B)\underline{P}(X|B) \leq \underline{P}(BX) \leq \bar{P}(B)\bar{P}(X|B) \leq \bar{P}(BX)$.
If $\bar{P}(X|B) \geq 0$ then $\underline{P}(BX) \leq \underline{P}(B)\bar{P}(X|B) \leq \bar{P}(BX) \leq \bar{P}(B)\bar{P}(X|B)$.

These properties can be derived from the coherence condition 6.3.2 by

6.4 THE GENERALIZED BAYES RULE

suitable choices of the gambles W , X , Y and Z . It will be shown in section 6.5 that avoiding sure loss and coherence can be characterized by a few of these properties, provided one of the domains \mathcal{H} and \mathcal{K} contains the other.

6.4 The generalized Bayes rule

Using the notation $G(X|B) = B(X - P(X|B))$ for the marginal gamble on X contingent on B , property 10 of section 6.3.5 states that $\underline{P}(G(X|B)) = 0$. This is a coherence relation between unconditional and conditional previsions. It will be called the **generalized Bayes rule** (GBR). In this section we discuss the implications of the GBR, especially for updating previsions. The reason for its importance is that, when B has positive lower probability, the conditional revision $\underline{P}(X|B)$ is uniquely determined by the unconditional revision \underline{P} through the GBR.

6.4.1 GBR theorem

Suppose that assumptions 6.3.1 hold, \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent, B is a set in the partition \mathcal{B} , and $\underline{P}(B)$ is defined and positive. If X is in \mathcal{H} and $G(X|B)$ is in \mathcal{K} , then $\underline{P}(X|B)$ is the unique value of μ such that $\underline{P}(B(X - \mu)) = 0$. Thus $\underline{P}(X|B)$ is uniquely determined by \underline{P} , through the generalized Bayes rule $\underline{P}(G(X|B)) = 0$.¹

Proof. By assumptions 6.3.1, $G(X|B) \in \mathcal{K}$. To show that $\underline{P}(G(X|B)) \geq 0$, set $Y = Z = G(X|B)$ in (a) of the coherence condition 6.3.2, and use property 6.2.6(j) to show that $\underline{P}(Y|\mathcal{B}) = 0$. Set $X = G(X|B)$ and $W = X$ in (b) of the coherence condition to give $\underline{P}(G(X|B)) \leq 0$. Coherence of \underline{P} implies that, when $\lambda > \mu$ and the previsions are defined, $\underline{P}(B(X - \mu)) \geq \underline{P}(B(X - \lambda)) + (\lambda - \mu)\underline{P}(B)$. Since $\underline{P}(B) > 0$, $\underline{P}(B(X - \mu))$ is a strictly decreasing function of μ . Hence $\underline{P}(X|B)$ is the unique solution of $\underline{P}(B(X - \mu)) = 0$. ◆

It follows simply from the interpretation of a contingent revision $\underline{P}(X|B)$, as the supremum price μ for which the contingent gamble $B(X - \mu)$ is desirable, that contingent previsions satisfy the GBR. The rule is especially important as an updating rule, for defining updated previsions in terms of initial previsions. In that case its justification relies on the updating principle, but not on the conglomerative principle.²

The GBR does indeed generalize Bayes' rule of conditioning. When the unconditional revision \underline{P} is linear, the GBR becomes $\underline{P}(BX) = \underline{P}(X|B)\underline{P}(B)$. If also $\underline{P}(B) > 0$, this determines a linear conditional revision $\underline{P}(X|B) = \underline{P}(BX)/\underline{P}(B)$. When $X = A$ is an event, this reduces to the familiar Bayes' rule $\underline{P}(A|B) = \underline{P}(A \cap B)/\underline{P}(B)$. The conditional revisions defined by the GBR can always be written as lower envelopes of linear conditional revisions, as follows.

6.4.2 Lower envelope theorem

Suppose that \underline{P} is coherent, $\underline{P}(B) > 0$, and $\underline{P}(X|B)$ satisfies the GBR. Define $\underline{P}(X|B)$ for linear previsions P by Bayes' rule, $P(X|B) = P(BX)/P(B)$. Then $\underline{P}(X|B) = \min\{\underline{P}(X|B) : P \in \mathcal{M}(\underline{P})\} = \min\{P(X|B) : P \in \text{ext } \mathcal{M}(\underline{P})\}$. On the domain where it is determined by the GBR, $\underline{P}(\cdot|B)$ is separately coherent.

Proof. Using the GBR, $0 = \underline{P}(G(X|B)) = \min\{\underline{P}(G(X|B)) : P \in \mathcal{M}(\underline{P})\} = \min\{\underline{P}(BX) - \underline{P}(X|B)\underline{P}(B) : P \in \mathcal{M}(\underline{P})\} = \min\{\underline{P}(B)(\underline{P}(X|B) - \underline{P}(X|B)) : P \in \mathcal{M}(\underline{P})\}$. Hence, since $\underline{P}(B) \geq \underline{P}(B) > 0$ when $P \in \mathcal{M}(\underline{P})$, $\min\{\underline{P}(X|B) : P \in \mathcal{M}(\underline{P})\} = \underline{P}(X|B)$. By Theorem 3.6.2(c), the minimum is achieved by some extreme point of $\mathcal{M}(\underline{P})$. It follows from linearity of P that $P(\cdot|B)$ is a linear prevision and $P(B|B) = 1$. By Corollary 2.8.6, $\underline{P}(\cdot|B)$ is coherent on its domain, and $\underline{P}(B|B) = 1$ by the GBR. ♦

Provided $\underline{P}(B) > 0$, the updated prevision $\underline{P}(\cdot|B)$ is the lower envelope of the updated class \mathcal{M}^B , which is obtained by conditioning the linear previsions in $\mathcal{M}(\underline{P})$ on B . The extreme points of \mathcal{M}^B form a subset of the conditioned extreme points of $\mathcal{M}(\underline{P})$. Often, the simplest way to apply the GBR is to apply Bayes' rule to the extreme points of $\mathcal{M}(\underline{P})$, and compute the lower envelope.

To illustrate this, we give two simple examples. In both examples, the conditional probabilities determined by the GBR are highly imprecise. More substantial examples will be given in section 6.6.

6.4.3 Two tosses of a coin

Suppose that a fair coin is tossed twice, in such a way that the second toss may depend on the outcome of the first, but You know nothing about the type or degree of dependence. The model \underline{P} suggested in Example 5.13.4 is the lower envelope of all additive probabilities P which assign $P(H_1) = P(H_2) = \frac{1}{2}$ and have an arbitrary degree of dependence between tosses. The two extreme points P_1 and P_2 of $\mathcal{M}(\underline{P})$ are characterized by $P_1(H_1 \cap H_2) = \frac{1}{2}$ and $P_2(H_1 \cap H_2) = 0$.

Now suppose that You observe the outcome of the first toss, represented by the partition $\mathcal{B} = \{H_1, T_1\}$. Because of the symmetry between heads and tails, it suffices to consider the observation H_1 . The extreme points P_1 and P_2 can be conditioned by Bayes' rule, giving $P_1(H_2|H_1) = P_1(H_1 \cap H_2)/P_1(H_1) = 1$, and $P_2(H_2|H_1) = P_2(H_1 \cap H_2)/P_2(H_1) = 0$. The updated lower prevision $\underline{P}(\cdot|H_1)$, obtained as the lower envelope of $P_1(\cdot|H_1)$ and $P_2(\cdot|H_1)$, is therefore vacuous. The GBR produces vacuous updated probabilities concerning the second toss, whatever the outcome of the first, even though Your initial probabilities $\underline{P}(H_2) = \bar{P}(H_2) = \frac{1}{2}$ were precise.

It may seem strange that precise probabilities can become vacuous when

6.4 THE GENERALIZED BAYES RULE

You obtain additional information.³ To see that it is reasonable, think of the extreme points P_1 and P_2 as hypotheses about how the experiment is performed. Suppose that the first toss is made in the usual way, but the second outcome is completely determined by the first: it is either identical (under hypothesis 1) or opposite (hypothesis 2). The model \underline{P} is reasonable if You are completely ignorant about which hypothesis is true. Before observing the first toss, Your probabilities concerning the second are precise because the two hypotheses imply the same probability $P(H_2) = \frac{1}{2}$. After observing the first toss, Your probabilities become vacuous because the two hypotheses predict different outcomes for the second toss. Observing the first toss introduces indeterminacy concerning the second toss, due to disagreement between the hypotheses.

This shows that receiving extra information can sometimes be a bad thing, in the sense that it is certain to produce indeterminacy and indecision. This effect is actually quite common in practice, especially when artificial randomization is involved. To illustrate, suppose that two treatments are compared by allocating them randomly to two different experimental units, and observing which unit responds better. (For example, two medical treatments might be tried on two people.) Consider the event A , that the unit given treatment 1 will respond better than the unit given treatment 2, and assume (as a null hypothesis) that the two treatments have identical effects. Before the random allocation of treatments, the event A has precise probability $\frac{1}{2}$. But after You learn which unit is allocated treatment 1, Your beliefs about A may be quite indeterminate, depending on what You know about the two units.⁴ The standard (frequentist) analysis of randomized experiments is based on the precise initial probabilities, and ignores the extra information about the outcome of the randomization.⁵

6.4.4 The three prisoners

The same effect is seen in the problem of the three prisoners, described in Example 5.13.10. Prisoner a learns from the governor either that b will be executed (event B), or that c will be executed (event C), so the partition is $\mathcal{B} = \{B, C\}$. The unconditional lower prevision \underline{P} , defined in 5.13.10, is the lower envelope of two linear previsions P_1 and P_2 , which are uniform distributions on the sets $\{ab, bc, cb\}$ and $\{ac, bc, cb\}$ respectively.

After observing $B = \{ab, cb\}$, P_1 and P_2 are updated to $P_1(\cdot|B)$, which is uniform on $\{ab, cb\}$, and $P_2(\cdot|B)$, which assigns probability one to $\{cb\}$. The event that a will be reprieved is $R = \{ab, ac\}$. Using the GBR, R has updated probabilities $\underline{P}(R|B) = P_2(R|B) = 0$ and $\bar{P}(R|B) = P_1(R|B) = \frac{1}{2}$. Similarly $\underline{P}(R|C) = 0$ and $\bar{P}(R|C) = \frac{1}{2}$. Whatever the governor tells him, prisoner a can conclude only that he will *probably* be executed.⁶

The updated probabilities of reprieve are again quite imprecise, although the initial probability was precisely $\frac{1}{3}$. The governor's answer introduces indeterminacy concerning R , because prisoner a is completely ignorant about how the governor would have chosen his answer if R were to occur.

6.4.5 Conditions for applying the GBR

Theorem 6.4.1 states that, under certain assumptions, conditional previsions are fully determined by unconditional ones through the GBR. This suggests that initial previsions should be updated automatically, on receiving new information, simply by applying the rule. Some qualifications are needed however, because the assumptions made in 6.4.1 are stronger than they may appear.

The first assumption is that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent. Suppose that, in assessing these previsions, You begin by directly assessing \underline{P} . If $\underline{P}(B)$ is positive then conditional previsions $\underline{P}(X|B)$, to be coherent with \underline{P} , must be computed through the GBR. But You will often be able to make more precise assessments of $\underline{P}(X|B)$ by other means, e.g. through judgements of independence.⁷ (Indeed, the preceding examples show that conditional probabilities defined by the GBR may be highly imprecise, so that other assessment strategies may be needed.) The conditional assessments $\underline{P}(X|B)$ are then incoherent with \underline{P} , simply because they provide extra information about unconditional previsions that can be used to construct a more precise model \underline{P} .

To illustrate this in the three prisoners problem, suppose prisoner a decides to adopt precise updated probability $\frac{1}{3}$ for reprieve, equal to his initial probability, because he judges that the governor's answer is entirely irrelevant to (or independent of) his prospects for reprieve. Because these assessments disagree with those computed by the GBR, they are incoherent with the lower prevision \underline{P} defined in Example 5.13.10. The new assessments (together with those in 5.13.10) imply, by natural extension, precise unconditional probabilities $P(\{ab\}) = P(\{ac\}) = \frac{1}{6}$. (They are tantamount to judging that, when a is to be reprieved, the governor is equally likely to answer B or C .)

In general, You should make assessments of conditional and unconditional previsions that avoid sure loss, and compute their natural extensions as in section 8.1. The resulting models \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ will be coherent, and therefore satisfy the GBR, but it is not necessary that $\underline{P}(\cdot|\mathcal{B})$ be constructed by first assessing \underline{P} and then applying the rule. It is equally valid to construct \underline{P} from earlier assessments of conditional previsions.⁸

Use of the GBR for updating beliefs relies on a second assumption, that Your new information establishes only that B has occurred. If You learn

that a football game is a cup-tie for which $B = \{\text{Win}, \text{Lose}\}$ must occur, this provides extra information that may influence Your updated previsions. It is not valid to apply the GBR (without further assumptions), because the new information cannot be identified with observation of B . This issue is discussed in section 6.11.

A third assumption is that B has positive lower probability. When $\underline{P}(B) = 0$, the unconditional prevision \underline{P} provides no information about conditional previsions $\underline{P}(X|B)$, and the natural extension of \underline{P} is the vacuous conditional prevision $\underline{P}(X|B) = \inf\{X(\omega): \omega \in B\}$.⁹

The fourth assumption is that \underline{P} is defined on a sufficiently large domain \mathcal{K} . To determine $\underline{P}(X|B)$, \mathcal{K} must include at least some gambles of the form $B(X - \mu)$ for real values of μ . When $\underline{P}(B) > 0$ and A is an event, $\underline{P}(A|B)$ is uniquely determined by the values of $\underline{P}(A \cap B - \mu B)$ over all real μ , but it is not necessarily determined by the restriction of \underline{P} to events.¹⁰ In general, $\underline{P}(A|B)$ is related to the lower probability \underline{P} only through inequality constraints, as follows.

6.4.6 Theorem

Suppose that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent, A and B are events, $B \in \mathcal{B}$ and $\underline{P}(B) > 0$. The following inequalities hold whenever their terms are defined.¹¹

$$\begin{aligned} \underline{P}(A \cap B)/\bar{P}(B) &\leq \underline{P}(A \cap B)/(\underline{P}(A \cap B) + \bar{P}(A^c \cap B)) \leq \underline{P}(A|B) \\ &\leq \min\{\underline{P}(A \cap B)/\underline{P}(B), \bar{P}(A \cap B)/\bar{P}(B)\} \\ &\leq \max\{\underline{P}(A \cap B)/\underline{P}(B), \bar{P}(A \cap B)/\bar{P}(B)\} \\ &\leq \bar{P}(A|B) \leq \bar{P}(A \cap B)/(\bar{P}(A \cap B) + \underline{P}(A^c \cap B)) \leq \bar{P}(A \cap B)/\underline{P}(B). \end{aligned}$$

Proof. Write $\lambda = \underline{P}(A \cap B)/(\underline{P}(A \cap B) + \bar{P}(A^c \cap B))$, so $0 \leq \lambda \leq 1$. Then $\underline{P}(B(A - \lambda)) = \underline{P}((1 - \lambda)AB - \lambda A^c B) \geq (1 - \lambda)\underline{P}(A \cap B) - \lambda \bar{P}(A^c \cap B) = 0$, so $\underline{P}(A|B) \geq \lambda$ by property 11 of 6.3.5. Use $\bar{P}(B) \geq \underline{P}(A \cap B) + \bar{P}(A^c \cap B)$ to show $\lambda \geq \underline{P}(A \cap B)/\bar{P}(B)$. The last two inequalities can be obtained from these by writing $\bar{P}(A|B) = 1 - \underline{P}(A^c|B)$, and the other inequalities are simple consequences of property 14 of 6.3.5. ◆

6.5 Coherence axioms

In this section we show that coherence of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$, defined in 6.3.2, can be characterized in terms of simpler axioms in some special cases. We consider the cases where one domain contains the other, \mathcal{B} is finite, or either \underline{P} or $\underline{P}(\cdot|\mathcal{B})$ is a linear prevision. The main results are stated as theorems, but only outlines of the proofs are given. It is assumed throughout that assumptions 6.3.1 hold. The first result gives a general characterization of coherence.

6.5.1 General axioms

Suppose that assumptions 6.3.1 hold. Then \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ avoid sure loss if and only if they satisfy the axiom

$$(C4) \text{ if } X \in \mathcal{K}, Y \in \mathcal{H} \text{ and } X \leq Y \text{ then } \underline{P}(X) \leq \sup \bar{P}(Y|\mathcal{B}).$$

Let $\bar{P}_B(Y|\mathcal{B})$ denote the gamble $B\underline{P}(Y|B) + B^c\bar{P}(Y|\mathcal{B})$. Then \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent if and only if they satisfy the two axioms

$$(C5) \text{ if } X \in \mathcal{K}, Y \in \mathcal{H} \text{ and } X \geq Y \text{ then } \underline{P}(X) \geq \inf \underline{P}(Y|\mathcal{B})$$

$$(C6) \text{ if } B \in \mathcal{B}, X \in \mathcal{K}, Y \in \mathcal{H} \text{ and } X \leq Y \text{ then } \underline{P}(X) \leq \sup \bar{P}_B(Y|\mathcal{B}).$$

Proof. Replace Y by $-Y$ in Definition 6.3.2 to obtain C4. Assuming C4, it suffices to prove $\sup T \geq 0$ when $T = G(X) + G(V|\mathcal{B})$. Let $Y = \sup T + \underline{P}(X) - G(V|\mathcal{B}) = \sup T + X - T \geq X$. By separate coherence, $\bar{P}(Y|\mathcal{B}) = \sup T + \underline{P}(X)$. By C4, $\underline{P}(X) \leq \sup \bar{P}(Y|\mathcal{B}) = \sup T + \underline{P}(X)$, so $\sup T \geq 0$.

Take $X = 0$, $Z = X$ in 6.3.2(a) to prove C5. Replace Y by $-B^cY$ and W by Y in 6.3.2(b) to get C6. To prove 6.3.2(a), it suffices to prove $\sup T \geq 0$ when $T = G(U) + G(V|\mathcal{B}) - G(Z)$. Let $X = Z - U$ and $Y = G(V|\mathcal{B}) + \underline{P}(Z) - \underline{P}(U) - \sup T$. Then $X - Y = \sup T - T \geq 0$, and $\underline{P}(Y|\mathcal{B}) = \underline{P}(Z) - \underline{P}(U) - \sup T$ by separate coherence. Apply C5 to show $\sup T \geq \underline{P}(Z) - \underline{P}(U) - \underline{P}(Z - U) \geq 0$, using coherence of \underline{P} .

To prove 6.3.2(b), it suffices to show $\sup T \geq 0$ when $T = G(X) + G(V|\mathcal{B}) - G(W|B)$. Let $Y = X - T + \sup T \geq X$. Use separate coherence to show that $\underline{P}(Y|B) \leq \underline{P}(X) + \sup T$, and $\bar{P}(Y|C) = \underline{P}(X) + \sup T$ when $C \neq B$, so $\bar{P}_B(Y|\mathcal{B}) \leq \underline{P}(X) + \sup T$. Apply C6 to give $\sup T \geq 0$. ◆

Axiom C5 formalizes the conglomerative principle and is equivalent to part (a) of the coherence condition 6.3.2, and C6 is equivalent to part (b). Under assumptions 6.3.1, each of C5 and C6 implies C4.¹ The three axioms hold automatically when all gambles in \mathcal{K} are \mathcal{B} -measurable. This means that a \mathcal{B} -marginal \underline{P} is coherent with any previsions conditional on \mathcal{B} .²

The three axioms simplify further when one of the domains contains the other.

6.5.2 Axioms when \mathcal{H} contains \mathcal{K} ³

Suppose \mathcal{H} contains \mathcal{K} . Then \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ avoid sure loss if and only if they satisfy

$$(C7) \underline{P}(X) \leq \sup \bar{P}(X|\mathcal{B}) \text{ whenever } X \in \mathcal{K}.$$

Coherence of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ is equivalent to the two axioms

$$(C8) \underline{P}(X) \geq \inf \underline{P}(X|\mathcal{B}) \text{ whenever } X \in \mathcal{K}$$

$$(C9) \underline{P}(X) \leq \sup \bar{P}_B(X|\mathcal{B}) \text{ whenever } X \in \mathcal{K} \text{ and } B \in \mathcal{B}.$$

6.5 COHERENCE AXIOMS

Proof. These axioms follow from those in Theorem 6.5.1 by taking $Y = X$. Conversely, use separate coherence to give $\underline{P}(X|B) \leq \underline{P}(Y|B)$ and $\bar{P}(X|B) \leq \bar{P}(Y|B)$ whenever $X \leq Y$. ◆

When the conditional previsions are interpreted as updated previsions, axiom C7 says that Your current buying price for X should be no larger than Your maximum updated selling price. (Otherwise, I can sell You X now and buy it back after observing B , forcing a sure loss.) C8 says that Your current buying price for X should be no smaller than Your minimum updated buying price. (I can force You to pay at least that much by waiting until B is observed.) C9 says that Your current buying price for X should be no larger than the maximum of Your updated buying price if B occurs and Your updated selling price if B fails. (Otherwise, by selling You X now and buying it back unless B occurs, I can force You to pay more than $\underline{P}(X|B)$ if B occurs, and to lose if B fails.)

6.5.3 Axioms when \mathcal{K} contains \mathcal{H}

Suppose \mathcal{K} contains \mathcal{H} . Then \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ avoid sure loss if and only if they satisfy

$$(C10) \bar{P}(G(X|\mathcal{B})) \geq 0 \text{ whenever } X \in \mathcal{K}.$$

Coherence of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ is equivalent to the two axioms

$$(C11) \underline{P}(G(X|\mathcal{B})) \geq 0 \text{ whenever } X \in \mathcal{K}$$

$$(C12) \underline{P}(G(X|B)) = 0 \text{ whenever } X \in \mathcal{K} \text{ and } B \in \mathcal{B}.$$

Proof. C4 implies C10 by taking $X = Y = -G(X|\mathcal{B})$ and using separate coherence. Conversely, if $X \leq Y$ then $\underline{P}(X) \leq \underline{P}(Y) \leq \bar{P}(\bar{P}(Y|\mathcal{B}) - Y) = \bar{P}(\bar{P}(Y|\mathcal{B})) - \bar{P}(G(-Y|\mathcal{B})) \leq \sup \bar{P}(Y|\mathcal{B})$, using C10 and coherence of \underline{P} , so C4 holds.

C5 implies C11 by taking $X = Y = G(X|\mathcal{B})$, and C11 implies C5 using $\underline{P}(Y) \geq \underline{P}(G(Y|\mathcal{B})) + \underline{P}(\underline{P}(Y|\mathcal{B})) \geq \inf \underline{P}(Y|\mathcal{B})$. Thus C5 is equivalent to C11.

C6 implies $\underline{P}(G(X|B)) \leq 0$, by taking $X = Y = G(X|B)$, and C11 gives $\underline{P}(G(X|B)) = \underline{P}(G(BX|\mathcal{B})) \geq 0$. Thus C5 and C6 imply C12. It remains to prove that C11 and C12 imply C6. Since $Y - \bar{P}_B(Y|\mathcal{B}) = G(Y|B) - G(-B^cY|\mathcal{B})$, $\underline{P}(Y - \bar{P}_B(Y|\mathcal{B})) \leq \underline{P}(G(Y|B)) - \underline{P}(G(-B^cY|\mathcal{B})) \leq 0$, using C11 and C12. Hence $\underline{P}(Y) \leq \underline{P}(Y - \bar{P}_B(Y|\mathcal{B})) + \bar{P}(\bar{P}_B(Y|\mathcal{B})) \leq \sup \bar{P}_B(Y|\mathcal{B})$, from which C6 follows. ◆

Each of the last three axioms has a simple interpretation. Recall that $G(X|\mathcal{B})$ can be regarded as a two-stage gamble in which You observe B and pay $\underline{P}(X|B)$, then observe $\omega \in B$ and receive $X(\omega)$. Axiom C10 says that the two-stage gamble should not be strictly undesirable. (Otherwise, You will

pay me to accept $G(X|\mathcal{B})$ now, and I can sell You X for up to $\underline{P}(X|B)$ after observing B , forcing a sure loss.) C11 says that the two-stage gamble should be almost desirable.⁵ (It is almost desirable after You make the observation, as You are willing to pay up to $\underline{P}(X|B)$ for X .) Axiom C12, the generalized Bayes rule, says that the contingent gamble $G(X|B)$ should be only marginally desirable. (If it is strictly desirable, You are willing to pay more than $\underline{P}(X|B)$ for X contingent on B .)

When the conditional previsions are interpreted as updated previsions, C11 formalizes the conglomerative principle (6.3.3). It is equivalent to part (a) of the coherence condition 6.3.2, and also equivalent to each of the axioms $\underline{P}(X) \geq \underline{P}(\underline{P}(X|\mathcal{B}))$ and $\bar{P}(X) \leq \bar{P}(\bar{P}(X|\mathcal{B}))$. The GBR C12 formalizes the updating principle (6.1.6). C12 relates P to a single conditional prevision $\underline{P}(X|B)$, whereas the conglomerative axioms C5, C8 and C11 relate P to the collection of conditional previsions $\underline{P}(X|\mathcal{B})$. The next result shows that the conglomerative axioms are needed only when the partition \mathcal{B} is infinite. For finite \mathcal{B} , the GBR is sufficient for coherence.

6.5.4 Finite \mathcal{B}

Suppose \mathcal{B} is finite and \mathcal{H} contains \mathcal{H} . Then P and $P(\cdot|\mathcal{B})$ are coherent if and only if they satisfy the generalized Bayes rule (C12).⁶

Proof. Assume C12 holds, $\mathcal{B} = \{B_1, B_2, \dots, B_m\}$ and $X \in \mathcal{H}$. Then $G(X|\mathcal{B}) = \sum_{j=1}^m G(X|B_j)$, so $\underline{P}(G(X|\mathcal{B})) \geq \sum_{j=1}^m \underline{P}(G(X|B_j)) = 0$, using coherence of P . Thus C11 holds. The result follows from 6.5.3. ◆

Result 6.5.3 shows that, when \mathcal{H} contains \mathcal{H} , the conglomerative axiom C11 and the GBR C12 are necessary and sufficient for coherence. By 6.5.4, C12 is sufficient when \mathcal{B} is finite. In that case the conglomerative principle is not needed. On the other hand, when \mathcal{B} is uncountable it is usual that $P(B) = 0$ for all B in \mathcal{B} , and then $\underline{P}(BZ) \leq 0$ for all Z so that C12 follows from C11. In that case C11 (or C5 or C8) is sufficient for coherence, and the updating principle is not needed. When \mathcal{B} is countably infinite, neither C11 nor C12 is sufficient for coherence.

6.5.5 Linear conditional previsions

Suppose that the conditional prevision $P(\cdot|\mathcal{B})$ is linear, i.e. $P(X + Y|\mathcal{B}) = P(X|\mathcal{B}) + P(Y|\mathcal{B})$ when X and Y are in \mathcal{H} .⁷ Then condition 6.3.2(b) reduces to avoiding sure loss. Hence coherence of P and $P(\cdot|\mathcal{B})$ is equivalent to 6.3.2(a), or to axiom C5, or to C8 (when \mathcal{H} contains \mathcal{H}), or to C11 (when \mathcal{H} contains \mathcal{H}). In this case, C11 is equivalent to each of the conditions $\underline{P}(G(X|\mathcal{B})) = 0$, $\bar{P}(G(X|\mathcal{B})) = 0$, $\underline{P}(X) = \underline{P}(\underline{P}(X|\mathcal{B}))$ and $\bar{P}(X) = \bar{P}(\bar{P}(X|\mathcal{B}))$.⁸

6.5 COHERENCE AXIOMS

6.5.6 Linear unconditional prevision

Next suppose that P is a linear prevision. Then the GBR reduces to Bayes' rule. To be coherent with P , $\underline{P}(\cdot|B)$ must be a linear prevision whenever $P(B) > 0$. Here 6.3.2(a) reduces to avoiding sure loss, so coherence of P and $\underline{P}(\cdot|\mathcal{B})$ is equivalent to 6.3.2(b), or to axiom C6, or to C9 (when \mathcal{H} contains \mathcal{H}), or to C11 plus Bayes' rule (when \mathcal{H} contains \mathcal{H}), or to Bayes' rule alone (when \mathcal{H} contains \mathcal{H} and \mathcal{B} is finite).

6.5.7 Linear conditional and unconditional previsions

Finally, suppose that both P and $P(\cdot|\mathcal{B})$ are linear previsions. Then coherence of P and $P(\cdot|\mathcal{B})$ is equivalent to avoiding sure loss, or to either (a) or (b) of 6.3.2, or to the general axiom

(C13) if $X \in \mathcal{H}$, $Y \in \mathcal{H}$ and $X \geq Y$ then $P(X) \geq \inf P(Y|\mathcal{B})$.

When \mathcal{H} contains \mathcal{H} , this simplifies to

(C14) $P(X) \geq \inf P(X|\mathcal{B})$ whenever $X \in \mathcal{H}$.⁹

Using 6.5.3, when \mathcal{H} contains \mathcal{H} , coherence is equivalent to the conglomerative axiom

(C15) $P(X) = P(P(X|\mathcal{B}))$ whenever $X \in \mathcal{H}$.¹⁰

This axiom says that Your current buying and selling price for any gamble X must equal Your current prevision of the updated price for X that You will later adopt, after observing an event from \mathcal{B} .¹¹

When \mathcal{B} is finite and \mathcal{H} contains \mathcal{H} , coherence is equivalent to Bayes' rule: $P(X|B) = P(BX)/P(B)$ whenever $X \in \mathcal{H}$, $B \in \mathcal{B}$ and $P(B) > 0$. This is true also if \mathcal{B} is countable rather than finite, and P is countably additive on \mathcal{B} .¹²

More generally, Bayes' rule is not sufficient for coherence; the stronger axiom C15 is needed. (This implies Bayes' rule, by replacing X by BX .) When P is not countably additive, there may be no conditional previsions that are coherent with P , because those defined by Bayes' rule do not satisfy C15. If there is $P(\cdot|\mathcal{B})$ that is coherent with a given P , it must be almost unique in the sense that there is P -probability zero that any two solutions $P_1(X|\mathcal{B})$ and $P_2(X|\mathcal{B})$ differ by more than a positive amount δ .¹³ If P is countably additive on \mathcal{B} , then $P_2(X|\mathcal{B})$ must be equal to $P_1(X|\mathcal{B})$ except on a set of P -probability zero.

Axiom C15 and Bayes' rule respectively formalize versions of the conglomerative and updating principles for the case of precise probabilities. When $P(\cdot|\mathcal{B})$ are interpreted as updated previsions, the justification of coherence relies, in general, on the conglomerative principle, but not on the updating principle. (When \mathcal{B} is finite, either principle suffices.)

6.5.8 Comparison with Kolmogorov's theory

When P and $P(\cdot|\mathcal{B})$ are linear and coherent, it follows from C15 and separate coherence that they satisfy $P(YX) = P(P(YX|\mathcal{B})) = P(YP(X|\mathcal{B}))$ for all \mathcal{B} -measurable gambles Y . This property is similar to the implicit definition of conditional probability in the standard theory of probability, due to Kolmogorov (1933), and suggests there is a close connection between coherence and the standard theory. However, there are important differences, which we will now outline, between the two theories.¹⁴

In the standard theory, conditional previsions (or expectations) $P(X|\mathcal{A})$ are defined with respect to a σ -field of events \mathcal{A} , rather than a partition \mathcal{B} . Given a linear prevision P (assumed to be countably additive), define $P(X|\mathcal{A})$ to be any \mathcal{A} -measurable gamble which satisfies **Kolmogorov's condition** $P(YP(X|\mathcal{A})) = P(YX)$ for all \mathcal{A} -measurable Y . This defines $P(X|\mathcal{A})$ only up to a set of P -probability zero: such $P(X|\mathcal{A})$ exist, and any two versions agree except on a set of probability zero.¹⁵

Measurability with respect to a σ -field \mathcal{A} (defined in section 3.2.1) is different, in general, from measurability with respect to a partition \mathcal{B} (section 6.2.5), but the two definitions agree when \mathcal{A} is the σ -field made up of all unions of sets in \mathcal{B} . This is denoted by $\mathcal{A}(\mathcal{B})$. Consider first the special case $\mathcal{A} = \mathcal{A}(\mathcal{B})$, in which the two approaches are closely related.

In this case, since $P(X|\mathcal{B})$ is $\mathcal{A}(\mathcal{B})$ -measurable and satisfies C15, the coherence conditions (6.3.1 and 6.3.2) are at least as strong as Kolmogorov's condition. They are strictly stronger in general, because Kolmogorov does not require that $P(\cdot|\mathcal{A}(\mathcal{B}))$ is separately coherent, or even that it is a linear prevision.¹⁶ That is, he does not require that $P(X|\mathcal{A}(\mathcal{B}))(\omega) \geq \inf\{X(\omega): \omega \in B\}$ when $\omega \in B$, nor that $P(X+Y|\mathcal{A}(\mathcal{B})) = P(X|\mathcal{A}(\mathcal{B})) + P(Y|\mathcal{A}(\mathcal{B}))$. These constraints do hold when all the conditioning events B have positive probability, because then $P(X|\mathcal{A}(\mathcal{B}))$ is uniquely determined by Bayes' rule (which implies the constraints). In that case, coherence is equivalent to Kolmogorov's condition, and both reduce to Bayes' rule. But when any conditioning event B has probability zero, Kolmogorov allows previsions conditional on B to be completely arbitrary.

6.5.9 Events with probability zero

This indeterminacy is glossed over in the Kolmogorov theory, on the grounds that events of probability zero can be ignored. But it is common, e.g. in statistical problems with a continuous sample space, for the unconditional prevision P to assign probability zero to all possible observations B . The indeterminacy cannot be ignored after You observe the event B , because You can then choose Your updated probabilities in a completely arbitrary way without violating Kolmogorov's condition.

The essence of the problem is that, when $P(B) = 0$, the unconditional prevision P contains no information about previsions conditional on B . That is illustrated by Borel's paradox (section 6.10.1) and other examples (section 8.6.5). The Kolmogorov approach, which attempts to define conditional previsions purely in terms of unconditional ones, therefore seems misguided. The conditional previsions are determined by the unconditional ones only when the conditioning event has positive probability.

This difficulty is recognized by Kolmogorov, who concludes that it is 'inadmissible' to consider probabilities conditional on an 'isolated' event B with probability zero; it is necessary to embed B in a specific partition \mathcal{B} .¹⁷ His conclusion seems to be mistaken. In our approach, it is legitimate and reasonable to consider probabilities conditional on any single event B , whether or not these are determined by unconditional probabilities.¹⁸ The coherence axioms, such as C15, should be regarded as consistency conditions relating conditional probabilities to unconditional ones, not as implicit definitions of conditional probabilities in terms of unconditional ones. Kolmogorov appears to confuse indeterminacy with meaninglessness because he regards conditional probability as a derived concept that can be defined (when it is meaningful) in terms of unconditional probability. In our approach, conditional probability is just as fundamental as unconditional probability.¹⁹

So far we have considered a special case of the standard theory, in which $\mathcal{A} = \mathcal{A}(\mathcal{B})$ contains arbitrary unions of sets in \mathcal{B} .²⁰ The σ -field $\mathcal{A}(\mathcal{B})$ consists of all sets A such that, whatever $B \in \mathcal{B}$ is observed, You will learn whether or not A occurred. Thus $\mathcal{A}(\mathcal{B})$ models what You learn from the observation. When the σ -field \mathcal{A} is not of the form $\mathcal{A}(\mathcal{B})$ for some partition \mathcal{B} , it is not clear what \mathcal{A} is supposed to represent.²¹ This matters because conditional previsions $P(X|\mathcal{A})$ depend strongly on \mathcal{A} , as illustrated by the next example.

6.5.10 Example

Let Ω be any uncountable set, and let P be any linear prevision that assigns probability zero to countable subsets. Let the σ -field \mathcal{A}_1 consist of all subsets of Ω which are countable or have countable complement. Then the \mathcal{A}_1 -measurable gambles are those which are constant except on a countable set. Since countable sets have zero probability, it follows that the constant conditional previsions $P(X|\mathcal{A}_1) = P(X)$ satisfy Kolmogorov's condition. Now let \mathcal{A}_2 contain all subsets of Ω . All gambles are \mathcal{A}_2 -measurable, so $P(X|\mathcal{A}_2) = X$ satisfies Kolmogorov's condition.

Each σ -field has the same atoms (the singleton sets), but the two conditional previsions are completely different. Writing $\mathcal{B} = \{\{\omega\}: \omega \in \Omega\}$

for the set of singletons, $\mathcal{A}_2 = \mathcal{A}(\mathcal{B})$ represents observation of the true state ω . It is not clear what kind of observation \mathcal{A}_1 is supposed to represent. Although \mathcal{A}_1 contains all the singleton sets $\{\omega\}$, it is absurd to adopt the constant conditional previsions $P(X|\mathcal{A}_1) = P(X)$ after observing ω . These conditional previsions are sensible only when the observation provides no information about ω .

6.6 Examples of conditional previsions

When $\underline{P}(B) > 0$, conditional previsions $\underline{P}(\cdot|B)$ can be constructed from unconditional ones by the GBR. If \mathcal{B} is infinite, it is necessary to check that the resulting $\underline{P}(\cdot|\mathcal{B})$ is coherent with \underline{P} , by verifying the conglomerative axiom C8 or C11.

Another useful method of constructing coherent pairs $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ is to define them to be the lower envelopes of a class Γ of coherent, linear pairs $P_\gamma, P_\gamma(\cdot|\mathcal{B})$. Here $P_\gamma(\cdot|\mathcal{B})$ can be defined from P_γ through Bayes' rule, or through the continuous version of Bayes' rule involving density functions (Theorem 6.10.4). Example 6.6.9 shows that not all coherent pairs $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ can be constructed in this way, as lower envelopes, although they can when \mathcal{B} is finite.

In this section the GBR is used to construct conditional previsions for some of the examples from section 2.9. In all the examples, \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are each defined for all gambles ($\mathcal{H} = \mathcal{K} = \mathcal{L}(\Omega)$), \underline{P} is coherent, and $\underline{P}(\cdot|\mathcal{B})$ is separately coherent. (So assumptions 6.3.1 are satisfied.)

6.6.1 Vacuous previsions

The vacuous lower prevision (section 2.9.1) is $\underline{P}(X) = \inf X$. The **vacuous conditional previsions** $\underline{P}(\cdot|B)$ are defined by $\underline{P}(X|B) = \inf\{X(\omega): \omega \in B\}$. This models complete ignorance about ω , except for the knowledge that ω belongs to the updated possibility space B . As might be expected, the vacuous conditional previsions are coherent with the vacuous unconditional prevision. (C8 holds with equality, from which C9 follows.)

By applying C8 to $X = BY + B^c \sup Y$, we see that, when \underline{P} is vacuous, $\underline{P}(\cdot|B)$ must also be vacuous. This means that if Your current beliefs are vacuous then Your current dispositions to update must also be vacuous. On the other hand, vacuous conditional previsions can be coherent with non-vacuous unconditional previsions. Any \underline{P} satisfying $\underline{P}(B) = 0$ for all $B \in \mathcal{B}$ is coherent with the vacuous $\underline{P}(\cdot|\mathcal{B})$.¹ (When \mathcal{B} is infinite, such \underline{P} can be linear, e.g. Examples 6.6.4 and 6.6.5.) In this case, there is essentially no information in \underline{P} to constrain $\underline{P}(\cdot|\mathcal{B})$.

6.6 EXAMPLES OF CONDITIONAL PREVISIONS

6.6.2 Linear–vacuous mixtures

Here $\underline{P}(X) = (1 - \delta)P_0(X) + \delta \inf X$, where P_0 is a linear prevision and $0 < \delta < 1$. The GBR uniquely determines the conditional previsions

$$\underline{P}(X|B) = [(1 - \delta)P_0(BX) + \delta \inf\{X(\omega): \omega \in B\}] / [(1 - \delta)P_0(B) + \delta].$$

(This also holds when $P_0(B) = 0 = \underline{P}(B)$, giving the vacuous conditional prevision.)

Provided \mathcal{B} is finite, \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent (by 6.5.4). To see that they must be coherent even when \mathcal{B} is infinite, note that $\underline{P}(X|B) \rightarrow \inf\{X(\omega): \omega \in B\}$ as $P_0(B) \rightarrow 0$. Given $\alpha > 0$, there is $\beta > 0$ such that $\underline{P}(X|B) \leq \inf\{X(\omega): \omega \in B\} + \alpha$ whenever $P_0(B) \leq \beta$. By finite additivity of P_0 , there are only finitely many events (B_1, \dots, B_m) in the partition with $P_0(B) > \beta$. Let A denote the union of all other events in the partition, so that $G(X|\mathcal{B})(\omega) \geq -\alpha$ when $\omega \in A$. Hence $\underline{P}(G(X|\mathcal{B})) = \underline{P}(\sum_{j=1}^m G(X|B_j) + AG(X|\mathcal{B})) \geq \sum_{j=1}^m \underline{P}(G(X|B_j)) - \alpha = -\alpha$, using the GBR. Since α is arbitrary, this establishes axiom C11. Thus \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are always coherent.

Writing $P_1(X) = P_0(X|B) = P_0(BX)/P_0(B)$ for the updated prevision determined by P_0 , and $\delta(B) = \delta / [(1 - \delta)P_0(B) + \delta]$, the updated lower prevision is $\underline{P}(X|B) = (1 - \delta(B))P_1(X) + \delta(B)\inf\{X(\omega): \omega \in B\}$. This is also a linear–vacuous mixture, on the reduced possibility space B . Thus the linear–vacuous form is preserved under conditioning. The updated degree of imprecision $\delta(B)$ is at least as large as the initial imprecision δ , and $\delta(B) \rightarrow 1$ as $P_0(B) \rightarrow 0$.

6.6.3 Constant odds-ratio models

The lower prevision $\underline{P}(X)$ is the unique solution of $(1 - \tau)P_0((X - \underline{P}(X))^+) = P_0((X - \underline{P}(X))^+)$, where P_0 is a linear prevision and $0 < \tau < 1$. To compute the conditional previsions, replace X by $G(X|B)$ and use the GBR $\underline{P}(G(X|B)) = 0$, giving $(1 - \tau)P_0(G(X|B)^+) = P_0(G(X|B)^-)$. Now $P_0(G(X|B)^+) = P_0((X - \underline{P}(X|B))^+|B)P_0(B)$, where $P_0(\cdot|B)$ is defined by Bayes' rule, and similarly for $G(X|B)^-$. Assuming that $P_0(B) > 0$, we obtain $(1 - \tau)P_0((X - \underline{P}(X|B))^+|B) = P_0((X - \underline{P}(X|B))^+|B)$. This has the same form as the initial equation, except that P_0 is replaced by $P_0(\cdot|B)$.

This shows that, on observing B with $P_0(B) > 0$, the constant odds-ratio model should be updated to a model of the same form, with P_0 replaced by $P_0(\cdot|B)$.² The degree of imprecision τ is unchanged, unlike the linear–vacuous model. The updated lower probabilities are

$$\underline{P}(A|B) = \frac{(1 - \tau)P_0(A|B)}{1 - \tau P_0(A|B)} = \frac{(1 - \tau)P_0(A \cap B)}{P_0(B) - \tau P_0(A \cap B)},$$

provided $P_0(B) > 0$.

More generally, suppose that $P_0(\cdot|\mathcal{B})$ is linear and coherent with P_0 , and each $\underline{P}(\cdot|B)$ is the constant odds-ratio model defined by $P_0(\cdot|B)$ and τ . Then $(1-\tau)P_0(G(X|\mathcal{B})^+|\mathcal{B}) = P_0(G(X|\mathcal{B})^-|\mathcal{B})$. Hence $(1-\tau)P_0(G(X|\mathcal{B})^+) = P_0(G(X|\mathcal{B})^-)$, using C15, which means that $\underline{P}(G(X|\mathcal{B})) = 0$. By 6.5.3, the constant odds-ratio models \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent.

6.6.4 Zero-one valued lower probabilities

Let \underline{P} be the natural extension of the zero-one valued lower probability $\underline{P}(A) = 1$ if $A \in \mathcal{A}$, $\underline{P}(A) = 0$ otherwise, where \mathcal{A} is a filter of sets. When $B \in \mathcal{B} \cap \mathcal{A}$, the GBR implies that $\underline{P}(\cdot|B)$ must agree with \underline{P} . Define $\underline{P}(\cdot|B)$ to be the vacuous conditional prevision when $\underline{P}(B) = 0$. To show that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent, verify axioms C8 and C9 (or C12).

For example, let Ω be a countable set and let \mathcal{A} be an ultrafilter containing all sets with finite complements.³ The corresponding linear prevision P has $P(A) = 0$ for all finite sets A . For any partition \mathcal{B} of Ω , $\underline{P}(\cdot|\mathcal{B})$ is coherent with P . In particular, if \mathcal{B} partitions Ω into finite sets then P is coherent with the vacuous conditional previsions, even though P is not countably additive on \mathcal{B} . This shows that countable additivity of a linear prevision is not necessary for coherence with some conditional previsions.

The following examples are also concerned with the connection between coherence and countable additivity.

6.6.5 'Uniform distribution' on the integers

Let Ω be the set of non-zero integers, and define \underline{P} by $\underline{P}(X) = \liminf_{n \rightarrow \infty} (1/2n) \sum_{j=1}^n (X(j) + X(-j))$. This is a kind of 'uniform distribution' on the integers. Suppose you observe $B_n = \{-n, n\}$, meaning that the integer has absolute value n , and $\mathcal{B} = \{B_n : n = 1, 2, \dots\}$. Then $\bar{P}(B_n) = 0$, so C12 is satisfied by any $\underline{P}(\cdot|\mathcal{B})$. Let $P(\cdot|B_n)$ be the uniform distribution on B_n , defined by $P(X|B_n) = (X(-n) + X(n))/2$. Verify that C8 holds, so $\underline{P}(\cdot|\mathcal{B})$ is coherent with \underline{P} .

Now consider any conditional previsions $\underline{P}(\cdot|\mathcal{B})$ that are dominated by $P(\cdot|\mathcal{B})$, i.e. such that $\underline{P}(X|\mathcal{B}) \leq P(X|\mathcal{B})$ for every X . Since C8 holds, \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent. Thus \underline{P} is coherent with a very wide range of conditional previsions, ranging from linear to vacuous.

Next consider any linear prevision P in $\mathcal{M}(P)$. Again, since C8 and C12 continue to hold, P is coherent with any $\underline{P}(\cdot|\mathcal{B})$ dominated by $P(\cdot|\mathcal{B})$. Even when P is linear, it can be coherent with a wide range of conditional previsions. Here P is coherent with linear conditional previsions $P(\cdot|\mathcal{B})$, although P is not countably additive on \mathcal{B} since $P(B_n) = 0$ for all n .⁴

6.6.6 A non-conglomerable linear prevision⁵

Here is an example of a linear prevision P that is not coherent with any conditional previsions. As in the previous example, let Ω be the non-zero integers and $B_n = \{-n, n\}$. Let $P = (P^+ + P^-)/2$, where P^+ is the countably additive distribution on the positive integers with $P^+(\{n\}) = 2^{-n}$, and P^- is any finitely additive probability that assigns probability one to the negative integers and probability zero to every finite subset.

Then $P(B_n) = P(\{n\}) > 0$, so $P(\cdot|B_n)$ is uniquely determined by Bayes' rule and $P(\{n\}|B_n) = 1$. But consider the set A of positive integers. Since $P(A) = \frac{1}{2}$ but $P(A|B_n) = 1$ for all n , axiom C14 fails. Thus P and $P(\cdot|\mathcal{B})$ incur sure loss.⁶

Because Bayes' rule is a consequence of C15 or C12, there is no linear conditional prevision $P(\cdot|\mathcal{B})$ such that P and $P(\cdot|\mathcal{B})$ avoid sure loss, and there is no non-linear conditional prevision that is coherent with P . This means that, although the model P alone is coherent, there is no coherent way to update P after observing a set in \mathcal{B} . This suggests that the linear prevision P may not be a sensible model, despite its coherence as an unconditional prevision. The problem arises because P^- , and hence P , is not countably additive. This will be discussed in section 6.8.5.

6.6.7 'Uniform distribution' on the positive integers

The same problem arises for the model suggested as a 'uniform distribution' on the positive integers, $\underline{P}(X) = \liminf_{n \rightarrow \infty} (1/n) \sum_{j=1}^n X(j)$. Consider the partition $\mathcal{B} = \{B_n : n \geq 1\}$, where B_n contains the odd integer $2n - 1$ and all even integers of the form $k2^n$ for odd k . Then $\underline{P}(B_n) = \bar{P}(B_n) = 2^{-(n+1)}$, and $\underline{P}(\cdot|B_n)$ is uniquely determined by the GBR.

Let A be the set of odd positive integers. Then $\underline{P}(A \cap B_n) = \bar{P}(A \cap B_n) = 0$ because $A \cap B_n$ is a singleton, and the GBR implies that $\underline{P}(A|B_n) = \bar{P}(A|B_n) = 0$ for all n . Since $\underline{P}(A) = \bar{P}(A) = \frac{1}{2}$, this violates C7. Thus \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ incur sure loss.

Again, there are no conditional previsions $\underline{P}(\cdot|\mathcal{B})$ that are coherent with the 'uniform distribution' \underline{P} . The same is true of all the linear previsions P in $\mathcal{M}(P)$, which are the models suggested by de Finetti as 'uniform distributions'.⁷ (The same argument applies because all the probabilities were precise.) Although they are coherent as unconditional previsions, the models \underline{P} and P cannot be coherently updated.

6.6.8 Linear-vacuous mixtures

The linear previsions P in the two previous examples are unreasonable because they cannot be coherently updated. However, there are arbitrarily

small neighbourhoods of these linear previsions that can be coherently updated. Specifically, take \underline{P} to be the linear–vacuous mixture defined by P and some positive δ . It was shown in Example 6.6.2 that the conditional previsions $\underline{P}(\cdot|\mathcal{B})$ defined by the GBR are coherent with \underline{P} . (The proof used only finite additivity of P .) For any linear prevision that cannot be coherently updated, there is a slightly less precise model that can.⁸ The reason is that most (all but finitely many) of the updated previsions $\underline{P}(\cdot|B)$ are much less precise than the corresponding $P(\cdot|B)$, even when δ is very small, because the updated imprecision $\delta(B) \rightarrow 1$ as $P(B) \rightarrow 0$.⁹

6.6.9 Coherent models that are not lower envelopes

Here is an example of a coherent pair $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ which is not a lower envelope of coherent linear pairs $P, P(\cdot|\mathcal{B})$. (In fact, it is not dominated by any linear pair that avoids sure loss.) Let Ω and \mathcal{B} be as in Example 6.6.6, and let P_1 denote the non-conglomerable linear prevision (P) defined there. Construct P_2 in a similar way so that $P_2(\{n\}) = 0$ and $P_2(\{-n\}) = 3^{-n}$ for $n \geq 1$, and $P_2(A) = \frac{1}{2}$ where A denotes the set of positive integers. Let \underline{P} be the lower envelope of P_1 and P_2 . Then $\underline{P}(B_n) > 0$, so that $\underline{P}(\cdot|\mathcal{B})$ is uniquely determined by the GBR. By Theorem 6.4.2, $\underline{P}(X|B_n) = \min\{P_1(X|B_n), P_2(X|B_n)\} = \min\{X(n), X(-n)\}$. Thus $\underline{P}(\cdot|\mathcal{B})$ is the vacuous conditional prevision, and this is coherent with \underline{P} since C8 holds.

Now let P be any linear prevision in $\mathcal{M}(P)$. Linear conditional previsions $P(\cdot|\mathcal{B})$ are determined by P through Bayes' rule, since $P(B_n) \geq \underline{P}(B_n) > 0$ for all n . We will show that P and $P(\cdot|\mathcal{B})$ incur sure loss. Since $\mathcal{M}(\underline{P})$ is the convex hull of $\{P_1, P_2\}$, $P = \lambda P_1 + (1 - \lambda)P_2$ for some $0 \leq \lambda \leq 1$. When $\lambda = 0$, $P(A|B_n) = P_2(A|B_n) = 0$ for all n , whereas $P(A) = P_2(A) = \frac{1}{2}$, incurring sure loss. Suppose that $\lambda > 0$. As $n \rightarrow \infty$, $P(\{-n\})/P(\{n\}) = 2\lambda^{-1}(1 - \lambda)(\frac{2}{3})^n \rightarrow 0$ so that $P(\{n\}|B_n) \rightarrow 1$. Choose N sufficiently large so that $P(\{n\}|B_n) \geq 1 - \lambda/4$ for $n \geq N$, and let C be the set of non-zero integers greater than $-N$. Then $P_1(C) = \frac{1}{2}$ and $P(C) = \lambda P_1(C) + (1 - \lambda)P_2(C) \leq \lambda/2 + (1 - \lambda) = 1 - \lambda/2$. But $P(C|B_n) \geq 1 - \lambda/4$ when $n \geq N$, and $P(C|B_n) = 1$ when $1 \leq n < N$. Again this incurs sure loss.

Here \underline{P} can be coherently updated on observing B_n , to the vacuous conditional previsions, but none of the linear previsions in $\mathcal{M}(\underline{P})$ can be coherently updated. The lower previsions \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent, but they are not dominated by any linear pairs P and $P(\cdot|\mathcal{B})$ that are coherent. Compare with the lower envelope theorem 3.3.3, which shows that all coherent unconditional previsions \underline{P} are lower envelopes of linear previsions. The sensitivity analysis interpretation of \underline{P} , which is supported by the earlier theorem, breaks down in this example, and again in the next example.

6.6.10 A linear prevision with vacuous updated previsions

Dubins (1975) has constructed a linear prevision P that is coherent with the vacuous conditional previsions, but not with any linear conditional previsions. Consider a joint experiment in which $\omega = (m, n)$ represents observation of two positive integers m and n , and the partitions $\mathcal{B} = \{B_m : m \geq 1\}$, $\mathcal{C} = \{C_n : n \geq 1\}$ represent observation of m, n respectively. Let Q be a linear prevision on the positive integers that is zero–one valued on events and zero for all finite sets. (These exist by Theorem 3.6.8.) Define P on $\mathcal{L}(\Omega)$ to have its \mathcal{C} -marginal and previsions conditional on C_n all equal to Q . Writing $X^n(m) = X(m, n)$ and using Theorem 6.7.2, P is uniquely determined by $P(X) = P(P(X|\mathcal{C})) = Q(Q(X^n))$. Then m and n each have marginal prevision Q , and they are ‘independent’ in the sense that the prevision for m is unchanged after observing n .

Let A be any subset of Ω . Then $P(A) = Q(Q(A^n))$. Since A^n and $Q(A^n)$ take only the values zero and one, and $Q(A^n) = 0$ whenever A is finite, the same is true of P . Thus P is zero–one valued on events and zero for all finite sets. Also $P(B_m) = Q(Q(\{m\})) = 0$ for all m . By 6.6.4, P is coherent with the vacuous $\underline{P}(\cdot|\mathcal{B})$.¹⁰

Next we show that no linear $P(\cdot|\mathcal{B})$ is coherent with P . Suppose otherwise. Let $A = \{(m, n) : 1 \leq n \leq m\}$, the event that the second integer is no larger than the first. For m such that $P(A|B_m) \geq \frac{1}{2}$, define $j(m)$ to be the smallest integer j such that $P(\{(m, n) : 1 \leq n \leq j\}|B_m) \geq \frac{1}{2}$. Otherwise define $j(m) = m$. Partition A into the three sets

$$\begin{aligned} A_1 &= \{(m, n) : 1 \leq n < j(m), m \geq 1\}, \\ A_2 &= \{(m, j(m)) : m \geq 1\}, \\ A_3 &= \{(m, n) : j(m) < n \leq m, m \geq 1\}. \end{aligned}$$

Now $P(A_1|B_m) < \frac{1}{2}$ and $P(A_3|B_m) \leq \frac{1}{2}$ for all m . Hence $P(A_1) = P(A_3) = 0$, by the conglomerative axiom C15 and the fact that P is zero–one valued. Also $P(A_2) = Q(Q(A_2^n)) = 0$, because the sets $A_2^n = \{m : j(m) = n\}$ partition the positive integers and $Q(A_2^n)$ can be non-zero for at most one n . Hence $P(A) = P(A_1) + P(A_2) + P(A_3) = 0$. But $A^n = \{m : m \geq n\}$, so $Q(A^n) = 1$ for all n , and $P(A) = Q(Q(A^n)) = 1$. This is a contradiction, so there is no linear $P(\cdot|\mathcal{B})$ that is coherent with P .

6.7 Extension of conditional and marginal previsions

We now turn to the problem of natural extension. Suppose that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are specified on domains \mathcal{K} and \mathcal{H} , and are coherent. The general problem is to find their minimal coherent extensions, \underline{E} and $\underline{E}(\cdot|\mathcal{B})$, to the domain \mathcal{L} containing all gambles. The minimal coherent extensions

summarize the implications of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ for other previsions, both conditional and unconditional.

A general theory of natural extension will be developed in section 8.1. In this chapter we consider two special types of extension. The first problem, discussed in this section, is to construct coherent extensions of \mathcal{B} -marginal previsions and previsions conditional on \mathcal{B} . That is always possible. The second problem is to find conditional previsions $\underline{E}(\cdot|\mathcal{B})$ that are coherent with unconditional ones. That is always possible when \mathcal{B} is finite, but not when \mathcal{B} is infinite, as seen in examples 6.6.6 and 6.6.7. For these two problems we will characterize the minimal coherent extensions, and show in section 8.1 that they agree with the ‘natural extension’ when this is suitably defined.

6.7.1 \mathcal{B} -marginals

Suppose that the domain \mathcal{K} of \underline{P} contains only \mathcal{B} -measurable gambles. In this case, \underline{P} is called a **\mathcal{B} -marginal**, because it represents beliefs about which event B in \mathcal{B} will be observed.

A useful strategy for assessing upper and lower previsions $\bar{P}(X)$ and $\underline{P}(X)$ is to assess previsions $\bar{P}(X|B)$ and $\underline{P}(X|B)$ conditional on each set B in a partition \mathcal{B} , and also assess a \mathcal{B} -marginal \underline{P} .¹ The natural extensions of these assessments, given by the following theorem, are $\underline{E}(X) = \underline{P}(\underline{P}(X|\mathcal{B}))$ and $\bar{E}(X) = \bar{P}(\bar{P}(X|\mathcal{B}))$. When all the assessments are precise we obtain $E(X) = P(P(X|\mathcal{B}))$.

This assessment strategy is often useful in statistical problems, for constructing marginal probabilities concerning the statistical observation x from sampling models $P(\cdot|\Theta)$ and a Θ -marginal \underline{P} , which represents prior beliefs about the statistical parameter θ . (In this case $\Omega = \Theta \times \mathcal{X}$, and \mathcal{B} corresponds to the parameter space Θ .) The Θ -marginal can be extended to prior probabilities for (θ, x) jointly, and these determine the \mathcal{X} -marginal.²

The main result of this section characterizes the minimal coherent extensions of conditional and marginal previsions which are defined on arbitrary domains.

6.7.2 Marginal extension theorem

Suppose that \underline{P} is coherent on \mathcal{K} all gambles in \mathcal{K} are \mathcal{B} -measurable, and $\underline{P}(\cdot|\mathcal{B})$ is separately coherent on an arbitrary domain \mathcal{K} . Then there are extensions of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ to \mathcal{L} that are coherent. The minimal coherent extensions are \underline{E} and $\underline{E}(\cdot|\mathcal{B})$, where each $\underline{E}(\cdot|B)$ is the natural extension of $\underline{P}(\cdot|B)$ to \mathcal{L} , \underline{E} is defined for all \mathcal{B} -measurable gambles as the natural extension of \underline{P} , and \underline{E} is extended to \mathcal{L} by $\underline{E}(X) = \underline{E}(\underline{E}(X|\mathcal{B}))$.³ The minimal

6.7 EXTENSION OF CONDITIONAL AND MARGINAL PREVISIONS 315

coherent extension satisfies $\underline{E}(X) = \underline{P}(\underline{P}(X|\mathcal{B}))$ whenever the right-hand side is defined.

Proof. Because $\underline{P}(\cdot|\mathcal{B})$ is separately coherent, so is $\underline{E}(\cdot|\mathcal{B})$. Using this and coherence of \underline{E} on $\mathcal{G}(\mathcal{B})$ (the set of \mathcal{B} -measurable gambles), the axioms in section 2.3.3 can be verified to show that \underline{E} is coherent on \mathcal{L} . By separate coherence, $\underline{E}(X - \underline{E}(X|\mathcal{B})) = 0$ for any $X \in \mathcal{L}$, hence $\underline{E}(X - \underline{E}(X|\mathcal{B})) = 0$. Thus \underline{E} and $\underline{E}(\cdot|\mathcal{B})$ satisfy axioms C11 and C12 of 6.5.3 (plus assumptions 6.3.1), so they are coherent.

Clearly $\underline{E}(\cdot|\mathcal{B})$ is an extension of $\underline{P}(\cdot|\mathcal{B})$. When $X \in \mathcal{K}$, $\underline{E}(X|\mathcal{B}) = X$ by separate coherence, so $\underline{E}(X) = \underline{P}(X)$. Thus \underline{E} is an extension of \underline{P} . By the results of section 3.1.2, $\underline{E}(\cdot|\mathcal{B})$ is the minimal extension of $\underline{P}(\cdot|\mathcal{B})$ that is even separately coherent, and \underline{E} is the minimal coherent extension of \underline{P} to $\mathcal{G}(\mathcal{B})$. Let \underline{E}' and $\underline{E}'(\cdot|\mathcal{B})$ be any other coherent extensions of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ to \mathcal{L} . By property (5) of 6.3.5, $\underline{E}(X) \geq \underline{E}'(\underline{E}(X|\mathcal{B})) \geq \underline{E}(\underline{E}(X|\mathcal{B})) = \underline{E}(X)$, for all $X \in \mathcal{L}$. Thus \underline{E} and $\underline{E}(\cdot|\mathcal{B})$ are the minimal coherent extensions. ◆

The minimal coherent extension summarizes the implications of the assessments \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ for other gambles. The value $\underline{E}(X) = \underline{P}(\underline{P}(X|\mathcal{B}))$ is the supremum price You are committed to pay for X . You are committed to pay up to $\underline{P}(X|\mathcal{B})$ for X after observing B , and to pay up to $\underline{P}(\underline{P}(X|\mathcal{B}))$ now for $\underline{P}(X|\mathcal{B})$, and the two transactions have the effect of paying up to $\underline{E}(X)$ for X .

The theorem implies that a coherent \mathcal{B} -marginal prevision \underline{P} and separately coherent conditional previsions $\underline{P}(\cdot|\mathcal{B})$ are automatically coherent with each other.⁴ These previsions are not linked in any way through the coherence conditions. That is why they are separately extended in the theorem, by natural extension. It is only when \underline{P} is extended beyond the \mathcal{B} -measurable gambles that it is linked to the conditional previsions, through the conglomerative condition $\underline{P}(X) \geq \underline{P}(\underline{P}(X|\mathcal{B}))$.

In the special case where no \mathcal{B} -marginal is specified, \mathcal{K} is empty and \underline{E} is vacuous on $\mathcal{G}(\mathcal{B})$. The minimal coherent extension to \mathcal{L} is then $\underline{E}(X) = \inf \underline{E}(X|\mathcal{B})$, with $\underline{E}(X) = \inf \underline{P}(X|\mathcal{B})$ and $\bar{E}(X) = \sup \bar{P}(X|\mathcal{B})$ if $X \in \mathcal{K}$.

6.7.3 Uniqueness

In general, the minimal coherent extension \underline{E} is not the only extension of \underline{P} that is coherent with $\underline{P}(\cdot|\mathcal{B})$.⁵ However, the extension $\underline{E}(X) = \underline{P}(\underline{P}(X|\mathcal{B}))$ is uniquely coherent whenever $\underline{P}(X|\mathcal{B})$ is precise and belongs to \mathcal{K} .⁶ If $\underline{P}(\cdot|\mathcal{B})$ is a linear prevision on \mathcal{L} and \underline{P} is defined on all \mathcal{B} -measurable gambles, then \underline{E} is the unique extension of \underline{P} that is coherent with $\underline{P}(\cdot|\mathcal{B})$. In statistical problems the sampling models $P(\cdot|\Theta)$ are commonly linear,

and then the predictive probabilities (\mathcal{X} -marginal) are uniquely determined by the prior probabilities concerning Θ (Θ -marginal).

For example, $\underline{E}(X) = \inf P(X|\mathcal{B})$ is the unique lower prevision that is coherent with the linear conditional previsions $P(\cdot|\mathcal{B})$ and has vacuous \mathcal{B} -marginal. The linear prevision $E(X) = P(P(X|\mathcal{B}))$ is the unique lower prevision that is coherent with linear $P(\cdot|\mathcal{B})$ and has linear \mathcal{B} -marginal P .

The minimal coherent extensions can be expressed as lower envelopes of coherent pairs of linear previsions, by the following result. (Compare with Theorem 3.4.1.)

6.7.4 Lower envelope theorem⁷

Under the assumptions of Theorem 6.7.2, the minimal coherent extensions $\underline{E}, \underline{E}(\cdot|\mathcal{B})$ are the lower envelopes of the class of all coherent linear pairs $P, P(\cdot|\mathcal{B})$ such that $P \in \mathcal{M}(\underline{P})$ and $P(\cdot|B) \in \mathcal{M}(P(\cdot|B))$ for all $B \in \mathcal{B}$.

Proof. Suppose that $P, P(\cdot|\mathcal{B})$ are such a pair. By Theorem 3.4.1, $P(X) \geq \underline{E}(X)$ when $X \in \mathcal{G}(\mathcal{B})$, and $P(X|\mathcal{B}) \geq \underline{E}(X|\mathcal{B})$ when $X \in \mathcal{L}$. Using C15, $P(X) = P(P(X|\mathcal{B})) \geq \underline{E}(\underline{E}(X|\mathcal{B})) = \underline{E}(X)$. Thus $P(\cdot|\mathcal{B})$ dominates $\underline{E}(\cdot|\mathcal{B})$ and P dominates \underline{E} .

To show that $\underline{E}(X)$ and $\underline{E}(X|\mathcal{B})$ can be achieved by such pairs, apply 3.4.1 again. For each B in \mathcal{B} there is $P(\cdot|B) \in \mathcal{M}(P(\cdot|B))$ such that $P(X|B) = \underline{E}(X|B)$, and there is $P_0 \in \mathcal{M}(\underline{P})$ such that $P_0(\underline{E}(X|\mathcal{B})) = \underline{E}(\underline{E}(X|\mathcal{B}))$. Define P by $P(Y) = P_0(P(Y|\mathcal{B}))$. Verify that $P \in \mathcal{M}(\underline{P})$, P and $P(\cdot|\mathcal{B})$ are coherent (they satisfy C15), $P(X|\mathcal{B}) = \underline{E}(X|\mathcal{B})$ and $P(X) = \underline{E}(X)$. ◆

6.7.5 Computing the extension

The coherent linear pairs $P, P(\cdot|\mathcal{B})$ in the theorem are determined by $P(\cdot|\mathcal{B})$ and the \mathcal{B} -marginal P_0 of P , since $P(Y) = P_0(P(Y|\mathcal{B}))$. The extension \underline{E} can be computed (in principle) by finding all \mathcal{B} -marginals P_0 for extreme points of $\mathcal{M}(\underline{P})$, finding all extreme points $P(\cdot|B)$ of each $\mathcal{M}(P(\cdot|B))$, constructing P from each combination by $P(Y) = P_0(P(Y|\mathcal{B}))$, and taking \underline{E} to be the lower envelope of all such P .

The computation of $\underline{E}(X)$ becomes much simpler when X is in \mathcal{H} , because it depends on $P(\cdot|\mathcal{B})$ only through $P(X|\mathcal{B})$. It can then be computed by $\underline{E}(X) = \min \{P_0(\underline{P}(X|\mathcal{B})): P_0 \in \mathcal{M}_{\mathcal{B}}(\underline{P})\}$, where $\mathcal{M}_{\mathcal{B}}(\underline{P})$ is the class of linear \mathcal{B} -marginals that dominate \underline{P} . When the partition \mathcal{B} is finite, this becomes $\underline{E}(X) = \min \{\sum_{B \in \mathcal{B}} P(X|B)P_0(B): P_0 \in \mathcal{M}_{\mathcal{B}}(\underline{P})\}$. Again, only the extreme points of $\mathcal{M}_{\mathcal{B}}(\underline{P})$ are needed.

For example, in statistical problems it is common that the linear sampling models have probability densities $f(x|\theta)$, and the Θ -marginals in $\mathcal{M}_{\Theta}(\underline{P})$

6.8 CONGLOMERABILITY

have densities $h(\theta)$. Then $\underline{E}(X)$ is the lower envelope of $P_0(P(X|\Theta)) = \int h(\theta)(\int X(\theta, x)f(x|\theta)dx)d\theta$ over the class of densities h . Thus the unique coherent extension \underline{E} is essentially the lower envelope of the joint densities $h(\theta)f(x|\theta)$, and its \mathcal{X} -marginal is essentially the lower envelope of the marginal densities $h^*(x) = \int h(\theta)f(x|\theta)d\theta$.

6.8 Conglomerability

The rest of this chapter is concerned with the problem of extending an unconditional prevision \underline{P} to conditional previsions that are coherent with \underline{P} . This kind of extension is especially important because of its role in updating beliefs, discussed in sections 6.1 and 6.11. When coherent extensions exist, they can be simply characterized. If $\underline{P}(B)$ is positive then $\underline{P}(\cdot|B)$ is determined by \underline{P} through the generalized Bayes rule. If $\underline{P}(B)$ is zero then the minimal coherent $\underline{P}(\cdot|B)$ is the vacuous conditional prevision. As seen in examples 6.6.6 and 6.6.7, however, there may be no $\underline{P}(\cdot|B)$ that are coherent with \underline{P} . The problem then arises of characterizing those \underline{P} that can be coherently extended to conditional previsions. That will be done through two axioms of conglomerability, whose meaning and justification are discussed in this section. In the following sections we will clarify the relationship between conglomerability and countable additivity (section 6.9), discuss the construction of non-vacuous conditional previsions when $\underline{P}(B)$ is zero (section 6.10), and consider when the GBR should be used to update beliefs (section 6.11).

We will assume that a coherent, unconditional lower prevision \underline{P} is defined on \mathcal{L} , or at least on a domain that includes all gambles of interest. If \underline{P} is initially specified on a smaller domain, it can be coherently extended to \mathcal{L} by natural extension.

6.8.1 Definition

Suppose \underline{P} is a coherent lower prevision on domain $\mathcal{L}(\Omega)$. If \mathcal{B} is a partition of Ω , \underline{P} is called \mathcal{B} -conglomerable when it satisfies the axiom

(P7) if $X \in \mathcal{L}$ and B_1, B_2, \dots are distinct sets in \mathcal{B} such that $\underline{P}(B_n) > 0$ and $\underline{P}(B_n X) \geq 0$ for all $n \geq 1$, then $\underline{P}(\sum_{n=1}^{\infty} B_n X) \geq 0$.

Call \underline{P} fully conglomerable when it is \mathcal{B} -conglomerable for every partition \mathcal{B} of Ω . This holds if and only if \underline{P} satisfies the axiom¹

(P8) if $X \in \mathcal{L}$ and $\{B_n: n \geq 1\}$ is any countable partition of Ω such that $\underline{P}(B_n) > 0$ and $\underline{P}(B_n X) \geq 0$ for all $n \geq 1$, then $\underline{P}(X) \geq 0$.

These axioms assert that certain countable sums of almost-desirable

contingent gambles should be almost-desirable. The term ' \mathcal{B} -conglomerable' is appropriate because \mathcal{B} -conglomerability is necessary and sufficient for \underline{P} and its natural extension $\underline{E}(\cdot|\mathcal{B})$ to satisfy the conglomerative axiom C11. The following theorem shows, in fact, that \mathcal{B} -conglomerability of \underline{P} is necessary and sufficient for the existence of conditional previsions $\underline{P}(\cdot|\mathcal{B})$ that are coherent with \underline{P} .

6.8.2 Coherent updating theorem

Suppose \underline{P} is a coherent lower prevision on domain $\mathcal{L}(\Omega)$, and \mathcal{B} is a partition of Ω .

- (a) There are separately coherent conditional previsions $\underline{P}(\cdot|\mathcal{B})$ on \mathcal{L} that are coherent with \underline{P} if and only if \underline{P} is \mathcal{B} -conglomerable.
- (b) If \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent, where $\underline{P}(\cdot|\mathcal{B})$ is separately coherent on \mathcal{L} , then $\underline{P}(\cdot|B)$ is uniquely determined by the GBR, $\underline{P}(B(X - \underline{P}(X|B))) = 0$, whenever $\underline{P}(B) > 0$.
- (c) Define $\underline{E}(\cdot|B)$ by the GBR when $\underline{P}(B) > 0$, and let $\underline{E}(X|B) = \inf\{X(\omega): \omega \in B\}$, the vacuous conditional prevision, when $\underline{P}(B) = 0$. If \underline{P} is \mathcal{B} -conglomerable then $\underline{E}(\cdot|\mathcal{B})$ is the minimal conditional prevision that is coherent with \underline{P} .

Proof. (b) was proved in Theorem 6.4.1. By Theorem 6.4.2 and the remark following Theorem 6.2.7, $\underline{E}(\cdot|\mathcal{B})$ is separately coherent. Suppose \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent. By 6.5.3, coherence is equivalent to C11 plus the GBR. Now $\underline{P}(X|B) = \underline{E}(X|B)$ when $\underline{P}(B) > 0$, by (b), and $\underline{P}(X|B) \geq \underline{E}(X|B)$ when $\underline{P}(B) = 0$. Thus $\underline{P}(X|\mathcal{B}) \geq \underline{E}(X|\mathcal{B})$. Using C11, $\underline{P}(X - \underline{E}(X|\mathcal{B})) \geq \underline{P}(X - \underline{P}(X|\mathcal{B})) \geq 0$, so \underline{P} and $\underline{E}(\cdot|\mathcal{B})$ also satisfy C11. They satisfy the GBR when $\underline{P}(B) > 0$, by definition of $\underline{E}(\cdot|B)$, and also when $\underline{P}(B) = 0$, because then the GBR follows from C11. Thus \underline{P} and $\underline{E}(\cdot|\mathcal{B})$ are coherent.

This proves that \underline{P} is coherent with some $\underline{P}(\cdot|\mathcal{B})$ if and only if it is coherent with $\underline{E}(\cdot|\mathcal{B})$, and then $\underline{E}(\cdot|\mathcal{B})$ is the minimal conditional prevision that is coherent with \underline{P} . Also, \underline{P} and $\underline{E}(\cdot|\mathcal{B})$ are coherent if and only if they satisfy C11. To complete the proof we need to show that this is equivalent to P7.

Suppose first that P7 holds. By super-additivity of \underline{P} , no more than n members of a partition can have $\underline{P}(B) \geq n^{-1}$, hence $\mathcal{B}^+ = \{B \in \mathcal{B}: \underline{P}(B) > 0\}$ is countable. Write $G(X|B) = B(X - \underline{E}(X|B))$. Then $G(X|\mathcal{B}) = X - \underline{E}(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} G(X|B) \geq \sum_{B \in \mathcal{B}^+} G(X|B)$, since $G(X|B) \geq 0$ when $\underline{P}(B) = 0$. If \mathcal{B}^+ is a finite set, $\underline{P}(G(X|\mathcal{B})) \geq \sum_{B \in \mathcal{B}^+} \underline{P}(G(X|B)) = 0$, by coherence of \underline{P} and the GBR. If \mathcal{B}^+ is countably infinite, apply P7 to $G(X|\mathcal{B})$ to give $\underline{P}(G(X|\mathcal{B})) \geq \underline{P}(\sum_{B \in \mathcal{B}^+} G(X|B)) \geq 0$. Thus P7 implies C11.

Conversely, suppose \underline{P} and $\underline{E}(\cdot|\mathcal{B})$ satisfy C11, and the hypotheses of P7 hold. Let $Y = \sum_{n=1}^{\infty} B_n X$. Then $\underline{E}(Y|B_n) = \underline{E}(X|B_n) \geq 0$ by the GBR, since

6.8 CONGLOMERABILITY

$\underline{P}(B_n X) \geq 0$ and $\underline{P}(B_n) > 0$. Hence $\underline{E}(Y|\mathcal{B}) \geq 0$ and $\underline{P}(Y) \geq \underline{P}(Y - \underline{E}(Y|\mathcal{B})) \geq 0$ by C11. Thus P7 holds. ♦

The natural extension $\underline{E}(\cdot|\mathcal{B})$ represents the updated buying prices for gambles that are implied by \underline{P} through the GBR. When $\underline{P}(B) = 0$, $\underline{E}(\cdot|B)$ is vacuous because \underline{P} provides no information about updated previsions.² When \underline{P} is \mathcal{B} -conglomerable, $\underline{E}(\cdot|\mathcal{B})$ is coherent with \underline{P} , and otherwise there are no conditional previsions that are coherent with \underline{P} .

6.8.3 Finite and uncountable partitions

When \mathcal{B} is finite, \mathcal{B} -conglomerability (axiom P7) holds trivially, and every coherent \underline{P} can be coherently extended to previsions conditional on \mathcal{B} . This is essentially because the conglomerative axiom C11 is implied by the GBR when \mathcal{B} is finite (see section 6.5.4), so any $\underline{P}(\cdot|\mathcal{B})$ that satisfies the GBR is automatically coherent with \underline{P} . When Ω is finite, every coherent \underline{P} is fully conglomerable, and \underline{P} can always be coherently updated.

When $\underline{P}(B) = 0$ for all B in \mathcal{B} , which is usual when \mathcal{B} is uncountable, \mathcal{B} -conglomerability again holds trivially. (In this case the minimal coherent extension is the vacuous conditional prevision.) Anyway, only countably many sets in \mathcal{B} can have $\underline{P}(B) > 0$. In order to check P7 we can restrict attention to the countable partition consisting of these sets and the complementary set. So \mathcal{B} -conglomerability is an issue only for countably infinite partitions \mathcal{B} , and we concentrate on that case.

6.8.4 Arguments for conglomerability

A lower prevision \underline{P} is \mathcal{B} -conglomerable if and only if it can be coherently extended to a conditional prevision $\underline{P}(\cdot|\mathcal{B})$, and \underline{P} is fully conglomerable if and only if this can be done for every partition \mathcal{B} of Ω . The main argument for the two conglomerability conditions is that an unconditional prevision should be capable of coherent extension to conditional previsions. There are two versions of the argument, based on the need to (i) update \underline{P} after observing B , and (ii) justify \underline{P} through assessments of contingent previsions.

First, if \mathcal{B} represents an experiment that could be performed, but there is no coherent way of updating the initial previsions \underline{P} after learning the outcome of the experiment, then the model \underline{P} seems unreasonable. But for any partition \mathcal{B} , you can envisage such an experiment by imagining that you will be told (by someone who knows the true state ω) which set B of \mathcal{B} contains ω . For coherent updating to always be possible, \underline{P} must be fully conglomerable.

Second, it will often be reasonable to ask you to justify your assessments \underline{P} by extending them to include assessments of contingent previsions $\underline{P}(\cdot|\mathcal{B})$.

that are coherent with \underline{P} . Again, this should be possible for any partition \mathcal{B} , although it is especially desirable for those partitions that are related in a natural way to the evidence on which \underline{P} is based.

Of course, these arguments rely on our specific definition of coherence (6.3.2), which is based on the updating and conglomerative principles. If either of these principles is rejected, then conglomerability is no longer necessary to allow extensions to conditional previsions. For example, de Finetti (1972, 1974) rejects the conglomerative principle.³ He apparently accepts the updating principle, and he defines both contingent and updated previsions $P(\cdot|\mathcal{B})$ from a linear prevision P through Bayes' rule. These are always coherent, according to his weaker definition of coherence. However, when P is not \mathcal{B} -conglomerable but it is updated by Bayes' rule, the initial and updated previsions can always be exploited to yield a sure loss;⁴ see the following example, as well as Examples 6.6.6 and 6.6.7. Such models seem wholly unreasonable.

When \mathcal{B} is countable and P is not \mathcal{B} -conglomerable, the only way to avoid this kind of sure loss is to adopt non-linear updated previsions $\underline{P}(\cdot|\mathcal{B})$. To do so, it is necessary to give up Bayes' rule, the GBR, and the updating principle. It is then unclear how to define the updated previsions.⁵ In any case, both principles do seem reasonable.

The role of the updating and conglomerative principles in justifying full conglomerability can be seen more clearly by arguing directly for axiom P8. Suppose that $\{B_n : n \geq 1\}$ is a countable partition of Ω , $P(B_n) > 0$ and $P(B_n X) \geq 0$ for all $n \geq 1$, and $\delta > 0$. Then $\underline{P}(B_n(X + \delta)) \geq \underline{P}(B_n X) + \delta \underline{P}(B_n) > 0$, so $B_n(X + \delta)$ is currently desirable. By the updating principle, $X + \delta$ is B_n -desirable, for all $n \geq 1$. By the conglomerative principle, $X + \delta$ is desirable, so $\underline{P}(X + \delta) = \underline{P}(X) + \delta \geq 0$. Since δ is arbitrary, $\underline{P}(X) \geq 0$. This establishes P8.⁶

Alternatively, one can by-pass both principles by arguing that $X + \delta = \sum_{n=1}^{\infty} B_n(X + \delta)$ should be desirable because it is a countable sum of desirable gambles. This goes beyond the arguments used in Chapter 2, which involved only finite sums of desirable gambles, but the extension seems unexceptionable because the sum is absolutely convergent, uniformly bounded, and at most one of the terms $B_n(X + \delta)$ is non-zero since $\{B_n : n \geq 1\}$ is a partition. This argument does not rely on the partition \mathcal{B} representing a real experiment or a natural assessment strategy.

These four arguments support full conglomerability (axiom P8) as a further axiom of rationality.⁷ We will add P8 to the coherence axioms (P1–P3) studied in earlier chapters.⁸ Coherent lower previsions that are not fully conglomerable, such as the ‘uniform distribution’ on the positive integers (sections 2.9.5 and 6.6.7), are not reasonable models.⁹

6.8.5 Example of non-conglomerability

To support this conclusion, it is instructive to look more closely at the non-conglomerable linear prevision P that was defined in example 6.6.6, and which is also examined by de Finetti (1972, section 9.5.2).¹⁰ The model can be re-interpreted as follows.

Identify Ω (the set of non-zero integers) with the product space $\Theta \times \mathcal{B}$, by identifying the integer n with the pair $(\text{sign}(n), |n|)$. The set $B_m = \{-m, m\}$ in \mathcal{B} represents observation of the positive integer m . The parameter space $\Theta = \{-, +\}$ represents two statistical hypotheses about the distribution of m . Conditional on the hypothesis $+$, it has a geometric distribution $P^+(\{m\}) = 2^{-m}$, which is countably additive. Conditional on the hypothesis $-$, it has a finitely additive distribution that assigns probability zero to each positive integer.

Suppose that initially the two hypotheses are judged equally probable. Bayes' rule then implies that $P(+, m) = 2^{-(m+1)}$, $P(-, m) = 0$ and $P(+|m) = 1$ for all $m \geq 1$. Whatever value of m is observed, You will become certain that $+$ is the correct hypothesis. If I know that You will update Your probabilities by Bayes' rule then I can easily exploit them, simply by betting on $+$ now, and betting against it after observing m , when the odds are certain to improve. Although P and $P(\cdot|\mathcal{B})$ are ‘coherent’ according to de Finetti's definition (because they satisfy Bayes' rule), they lead to behaviour that incurs sure loss and is clearly irrational.

6.8.6 Modifications to achieve coherence

Let us consider several ways in which this model can be modified to achieve coherence.

1. In any practical problem, the countable partition \mathcal{B} is presumably an idealization of a finite set of possible observations.¹¹ When \mathcal{B} is taken to be finite, Bayes' rule is sufficient for coherence (6.5.4). For example, if the possible observations are the integers $1, 2, \dots, M$ plus the set $C = \{M+1, M+2, \dots\}$, then Bayes' rule yields the conditional probabilities $P(+|m) = 1$ for $1 \leq m \leq M$ and $P(+|C) = (1 + 2^M)^{-1}$, which are coherent with P .

If \mathcal{B} can always be assumed to be finite then conglomerability holds automatically and there is no issue. However, in statistical inference it is often useful (and arguably necessary) to consider infinite partitions, so it is important to take this case seriously. We need some assurance that, when we use infinite sample spaces or infinite parameter spaces, we will not reach absurd conclusions.

2. If we accept the infinite partition \mathcal{B} , we must modify the linear prevision P as it is not \mathcal{B} -conglomerable. Now P was constructed from the two conditional distributions P^+ and P^- , and the equal prior probabilities for the two hypotheses. Because $P(+|m) = 1$ for all m , one way to obtain coherence is to increase the prior probability of $+$ from $\frac{1}{2}$ to 1. But it seems strange that Your initial beliefs about the hypotheses should be determined by the prospect of observing m .

That becomes clearer if the true hypothesis is determined by a random mechanism, such as tossing a fair coin. If the coin lands ‘heads’ then I will report a positive integer m_1 that has the geometric distribution P^+ . If the coin lands ‘tails’ then I will report a positive integer m_2 , Your beliefs about which are modelled by the ‘uniform distribution’ P^- . After learning m , what are Your beliefs about the outcome of the coin toss? As before, You will assign probability one to ‘heads’ after learning m , and now the equal prior probabilities are indisputable. Moreover, the geometric distribution for m_1 is reasonable since it can be physically generated, by taking m_1 to be the number of tosses before the coin lands ‘tails’. So only the ‘uniform distribution’ P^- is open to objection.

3. Could You have evidence about m_2 that makes P^- a reasonable model?¹² De Finetti advocates P^- as a model for complete ignorance about m_2 . He argues that we can be completely ignorant about a state ω from a finite or bounded continuous space Ω , and model our ignorance through a uniform distribution, so it is natural to suppose that we can be completely ignorant about the value of a positive integer, and model this through some ‘uniform distribution’ which must assign probability zero to each integer. De Finetti is right to suppose that You may be completely ignorant about a positive integer, but wrong to suppose that linear previsions can adequately model complete ignorance. The non-conglomerability of the ‘uniform distribution’ casts further doubt on the adequacy of linear previsions, in addition to the arguments in section 5.5. Complete ignorance can be properly modelled by the vacuous previsions, which are fully conglomerable.

4. De Finetti suggests that the ‘uniform distribution’ is merely a convenient idealization of a uniform distribution on a finite set $\{1, 2, \dots, M\}$, where M is a very large integer. The finite uniform distribution is countably additive and therefore conglomerable. Bayes’ rule yields $P(+|m) = (1 + 2^m/M)^{-1}$ if $1 \leq m \leq M$, $P(+|m) = 1$ if $m > M$. When the observation m lies between $M/3$ and M (which is probable under hypothesis $-$), the hypothesis $+$ has very small posterior probability, as compared with probability one under the previous model. This shows that it can be highly misleading to use a ‘uniform distribution’ on the positive integers as an approximation to a uniform distribution on a large finite set.

Our conclusion is that non-conglomerable models such as P^- are unacceptable. In the next section we will show that \mathcal{B} -conglomerability fails because P^- is not countably additive on \mathcal{B} .

6.9 Countable additivity

In the previous section we argued that conglomerability is a necessary condition of rationality. Next we will clarify the relationship between conglomerability and countable additivity, and the extent to which countable additivity is a condition of rationality.

As noted in section 6.8.3, \mathcal{B} -conglomerability is not an issue for finite or uncountable partitions, so we will assume that $\mathcal{B} = \{B_1, B_2, \dots\}$ is countably infinite. Suppose that P is a linear prevision. We say that P is **countably additive** on \mathcal{B} when $\sum_{n=1}^{\infty} P(B_n) = 1$. Examples 6.6.4 and 6.6.5 show that P can be \mathcal{B} -conglomerable without being countably additive on \mathcal{B} . Indeed, that happens whenever $P(B) = 0$ for all B in \mathcal{B} . It can happen even if $P(B) > 0$ for all B . So countable additivity is not necessary for \mathcal{B} -conglomerability.¹ The following result shows that it is sufficient.

6.9.1 Theorem

If P is a linear prevision on \mathcal{L} that is countably additive on a countable partition \mathcal{B} then P is \mathcal{B} -conglomerable. In fact, there are linear conditional previsions $P(\cdot|\mathcal{B})$ that are coherent with P .

Proof. It follows from countable additivity on \mathcal{B} and boundedness of X that $P(X) = \sum_{n=1}^{\infty} P(B_n X)$. This implies axiom P7. Define $P(\cdot|B)$ by Bayes’ rule when $P(B) > 0$, and let $P(\cdot|\mathcal{B})$ be any separately coherent linear prevision when $P(B) = 0$. Then $P(X) = \sum_{n=1}^{\infty} P(B_n X) = \sum_{n=1}^{\infty} P(B_n)P(X|B_n) = \sum_{n=1}^{\infty} P(B_n P(X|\mathcal{B})) = P(P(X|\mathcal{B}))$. Since axiom C15 holds, P and $P(\cdot|\mathcal{B})$ are coherent. ◆

As a corollary, any linear prevision that is countably additive on the finest partition of a countable space Ω is fully conglomerable and can always be coherently updated to a linear prevision. It might be suspected that countable additivity is necessary for *full* conglomerability, but that is not so. The zero–one valued additive probability in example 6.6.4, defined on a countable space Ω , is fully conglomerable but not countably additive. (Axiom P8 holds trivially because at most one set in a partition can have positive probability.) This example is unusual, however, because P takes only finitely many different values on events. When that is not the case, Schervish, Seidenfeld and Kadane (1984) have shown that countable additivity is necessary for full conglomerability.² Hence the following result holds.

6.9.2 Theorem

Suppose that P is a linear prevision on \mathcal{L} , and $P(A)$ takes infinitely many different values as A ranges over all subsets of Ω . Then P is fully conglomerable if and only if it is countably additive on every countable partition of Ω .

When Ω is a countable set, there is a rich supply of countably additive (hence fully conglomerable) linear previsions, defined on all subsets of Ω . But when Ω is uncountable, relatively few linear previsions on $\mathcal{L}(\Omega)$ are fully conglomerable. In fact, there is no fully conglomerable linear prevision P on $\mathcal{L}(\Omega)$ that takes infinitely many different values on events and satisfies $P(\{\omega\})=0$ for all $\omega \in \Omega$.³ For example, there is no fully conglomerable linear extension of Lebesgue measure to all gambles on the unit interval.

Nevertheless, there is a rich supply of fully conglomerable lower previsions, whatever the cardinality of Ω . This will be illustrated for uncountable Ω by the example of Lebesgue measure (Example 6.9.6). For countable Ω , it follows from the fact that lower envelopes of conglomerable linear previsions are always conglomerable.⁴

6.9.3 Theorem

Let \underline{P} be the lower envelope of a class of coherent lower previsions \underline{P}_γ , all defined on \mathcal{L} . If all \underline{P}_γ are \mathcal{B} -conglomerable (or fully conglomerable) then so is \underline{P} .

Proof. Verify that \underline{P} must satisfy P7 or P8 if all \underline{P}_γ do. ◆

6.9.4 Corollary

Let \underline{P} be the lower envelope of a class of linear previsions P_γ , all defined on \mathcal{L} . If \mathcal{B} is a countable partition and each P_γ is countably additive on \mathcal{B} , then \underline{P} is \mathcal{B} -conglomerable. If each P_γ is countably additive on every countable partition of Ω , then \underline{P} is fully conglomerable.

When Ω is countable, a rich supply of conglomerable lower previsions can be constructed as lower envelopes of countably additive, linear previsions. However, there are \mathcal{B} -conglomerable lower previsions that are not dominated by any \mathcal{B} -conglomerable linear prevision (Example 6.6.9). Similarly, there are fully conglomerable lower previsions that are not dominated by any fully conglomerable linear prevision. That is illustrated by the next two examples. The first example allows Ω to be either countable or uncountable.

6.9 COUNTABLE ADDITIVITY

6.9.5 Linear–vacuous mixtures

Let Ω be any infinite set, and let P_0 be any linear prevision on $\mathcal{L}(\Omega)$ that takes infinitely many different values on events but is not countably additive. (For examples, see sections 2.9.5, 6.6.6 or 6.9.6.) Let \underline{P} be the linear–vacuous mixture $\underline{P}(X) = (1 - \delta)P_0(X) + \delta \inf X$, where $0 < \delta < 1$. By Example 6.6.2, \underline{P} is fully conglomerable. The linear previsions that dominate \underline{P} have the form $(1 - \delta)P_0 + \delta P$, for some linear prevision P . Verify that $(1 - \delta)P_0 + \delta P$ takes infinitely many values on events⁵ and is not countably additive. By Theorem 6.9.2, none of these dominating linear previsions is fully conglomerable.

For example, let Ω be the set of positive integers and let P_0 be any ‘uniform distribution’ on Ω (section 2.9.5). The P_0 -vacuous mixture \underline{P} can be coherently updated, to another linear–vacuous mixture, after an observation from any partition \mathcal{B} . But (as shown in 6.6.7) there are partitions \mathcal{B} for which P_0 cannot be coherently updated, and the same is true of every linear prevision $(1 - \delta)P_0 + \delta P$ that dominates \underline{P} .

6.9.6 Inner Lebesgue measure

Let Ω be the unit interval $[0, 1]$, and let \underline{P} be the Lebesgue lower prevision on $\mathcal{L}(\Omega)$, defined in section 2.9.6. By Example 3.2.7, \underline{P} is the natural extension of Lebesgue measure (or of inner Lebesgue measure) to all gambles. We now show that \underline{P} is fully conglomerable.

Suppose $\{B_n : n \geq 1\}$ is a countable partition of Ω such that $\underline{P}(B_n) > 0$ and $\underline{P}(B_n X) \geq 0$ for all $n \geq 1$. To establish axiom P8, we need to show that $\underline{P}(X) \geq 0$. Let \mathcal{K} be the class of Borel-measurable gambles, so that $P = \underline{P}$ is linear and countably additive on \mathcal{K} , and $\underline{P}(Z) = \sup\{P(Y) : Y \leq Z, Y \in \mathcal{K}\}$. Given $\delta > 0$, find $Y_n \in \mathcal{K}$ such that $Y_n \leq B_n X$ and $P(Y_n) \geq -\delta 2^{-n}$. For each $\omega \in \Omega$, $Y_n(\omega) \leq 0$ unless $\omega \in B_n$. Hence $\sum_{n=1}^{\infty} Y_n(\omega)$ converges, possibly to $-\infty$, and $\sum_{n=1}^{\infty} Y_n \leq \sum_{n=1}^{\infty} B_n X = X$. Let $Y = \max\{\sum_{n=1}^{\infty} Y_n, \inf X\}$. Then $\inf X \leq Y \leq X$, Y is bounded, and Y is Borel-measurable since all Y_n are. Thus $Y \in \mathcal{K}$. By Lebesgue’s dominated convergence theorem, $P(Y) \geq \sum_{n=1}^{\infty} P(Y_n) \geq -\delta$. Hence $\underline{P}(X) = \sup\{P(Y) : Y \leq X, Y \in \mathcal{K}\} \geq 0$, as required.⁶

By Theorem 3.4.2, the linear previsions in $\mathcal{M}(\underline{P})$ are just the linear extensions of Lebesgue measure to $\mathcal{L}(\Omega)$, and none of these is countably additive. By Theorem 6.9.2, no such linear prevision is fully conglomerable.⁷ Thus the Lebesgue lower prevision is fully conglomerable, but none of its dominating linear previsions is fully conglomerable.

More generally, suppose that Ω is uncountable and P is a countably additive probability measure, defined on a σ -field of subsets of Ω . It is not usually possible to extend P to a fully conglomerable linear prevision on $\mathcal{L}(\Omega)$. (See the remarks following 6.9.2.) But it is possible to extend P to a

fully conglomerable lower prevision on $\mathcal{L}(\Omega)$. Simply take this to be the natural extension of P , as in the case of Lebesgue measure.

6.9.7 De Finetti's arguments against countable additivity

The arguments in favour of full conglomerability in section 6.8.4, together with Theorem 6.9.2, produce a strong argument in favour of countable additivity. Before discussing its implications, we must reply to the arguments presented by de Finetti, against the requirement of countable additivity.⁸ His arguments can be summarized as follows:

1. 'We can define *uniform* probability distributions on finite or bounded continuous spaces, so we should be able to do so on countably infinite spaces.' But why are uniform distributions needed? De Finetti suggests that they model complete ignorance, but the arguments in sections 5.5, 6.6.7 and 6.8.6 indicate otherwise. De Finetti argues that a uniform distribution on a countable set is needed, for example, to model updated beliefs when Your initial distribution for ω is uniform on the unit interval, and You learn only that ω is a rational number. However, on our theory of conditioning (and apparently also on de Finetti's), the updated beliefs here are entirely indeterminate, because the set of rationals has initial probability zero. The updated probabilities defined by natural extension are vacuous, rather than uniform, on the set of rationals.
2. 'Lebesgue measure can be extended to finitely additive probabilities, defined on all subsets of the unit interval, but these are not countably additive.' If it could be taken for granted that probabilities should be finitely additive, then this would indeed be a strong argument against the additional requirement of countable additivity. But it should rather be seen as an argument against finite additivity (see section 6.9.8).
3. 'The limit of a convergent sequence of linear previsions is a linear prevision, but countable additivity may not be preserved under the limiting operation.'⁹ But why should the limit of a sequence of reasonable (countably additive) probability models be a reasonable model? What interpretation of the limiting process would justify this requirement?
4. 'No justification has been given for requiring countable additivity, except on grounds of mathematical convenience.' We agree with de Finetti that 'mathematical convenience' is an inadequate justification, but the argument for conglomerability is much stronger.

6.9.8 Conclusions

We accept full conglomerability as a rationality axiom. Non-conglomerability should be ruled out, not because its consequences are merely strange or

6.9 COUNTABLE ADDITIVITY

paradoxical,¹⁰ but because it leads to behaviour that is clearly irrational: it produces a sure loss (in the case of linear previsions) or inconsistency (in general). Linear previsions should therefore be countably additive.¹¹

But consider the implications. There is no countably additive extension of Lebesgue measure to all subsets of the unit interval. (This example is typical of uncountable spaces Ω .) So there is no coherent extension \underline{P} of Lebesgue measure that satisfies the three conditions: (a) \underline{P} is defined on all gambles, (b) \underline{P} is fully conglomerable, and (c) \underline{P} is a linear prevision. Condition (a) is a requirement that specified previsions can be coherently extended to any larger domain. Similarly, (b) requires that they can be coherently extended to conditional previsions, for any partition of Ω . It is reasonable to require that specified previsions be capable of coherent extension. So we regard both (a) and (b) as rationality requirements. But they are incompatible, in general, with (c). This is another argument against linearity.

Conditions (a) and (b) are compatible, because they are satisfied by the natural extension of Lebesgue measure to all gambles (section 6.9.6). This is surely a reasonable model (assuming that Lebesgue measure itself is reasonable), because it merely expresses the implications of Lebesgue measure for non-measurable gambles. But any linear extension of Lebesgue measure is unreasonable, because for some partition \mathcal{B} it cannot be coherently updated.¹²

Those who insist on additivity and linearity must reject either conglomerability, or the possibility of extensions to larger domains. De Finetti argues that coherent extension should always be possible, and he is therefore forced to give up conglomerability. Kolmogorov's theory of probability, which is based on an axiom of countable additivity, cannot admit extensions beyond a σ -field of sets and its corresponding space of measurable gambles. The two desiderata (conglomerability and arbitrary extension) can be reconciled quite easily, once additivity is dropped.

Finally, consider the implications of these results for the sensitivity analysis interpretation of lower previsions. All coherent lower previsions are lower envelopes of linear previsions. But not all coherent models are rational. If the rationality axioms are strengthened to include conglomerability as well as coherence, some rational lower previsions are not lower envelopes of rational linear previsions; see Examples 6.6.9, 6.9.5 and 6.9.6. Inner Lebesgue measure, for instance, is not a lower envelope of countably additive probability measures. Even when a linear prevision is fully conglomerable, there may be no linear conditional previsions $P(\cdot|\mathcal{B})$ that are coherent with it (6.6.10). In these examples the lower prevision \underline{P} can be coherently updated, but a sensitivity analyst would be unable, in general, to coherently update the linear previsions in $\mathcal{M}(P)$.¹³ This is a further argument against the sensitivity analysis interpretation and the dogma of ideal precision.

6.10 Conditioning on events of probability zero

An unconditional prevision \underline{P} uniquely determines $\underline{P}(\cdot|B)$ when $\underline{P}(B) > 0$, through the GBR. But when $\underline{P}(B) = 0$, the minimal coherent extension $\underline{E}(\cdot|B)$ is vacuous, and \underline{P} apparently provides no information about updated previsions.¹ This will be illustrated by Borel's paradox. Non-vacuous updated previsions $\underline{P}(\cdot|B)$ can be constructed from \underline{P} only if further information is provided. One way to do this is to model B as a limit of conditioning events with positive probability, and define $\underline{P}(\cdot|B)$ as the limit of the corresponding conditional previsions. Under further assumptions, this justifies a version of Bayes' rule for density functions.

6.10.1 Borel's paradox²

Let Ω represent the surface of a sphere, e.g. an idealization of the earth's surface. A point ω is chosen at random, according to a uniform distribution on Ω . The observation B is that ω lies on a particular **great circle**, i.e. a circle that divides the surface area into two equal halves. Conditional on B , what is the distribution of the point ω ?

To answer this, establish a coordinate system on Ω , by arbitrarily fixing two poles and a half-circle of zero longitude between the poles. Identify points ω by their coordinates (θ, ψ) , where θ is **latitude** ($-\pi/2 \leq \theta \leq \pi/2$) and ψ is **longitude** ($-\pi < \psi \leq \pi$). The uniform distribution on Ω means that θ and ψ are chosen independently, θ has density $\frac{1}{2}\cos\theta$, and ψ has uniform density $1/2\pi$.³

Consider two ways of fixing these coordinates:

1. Choose the half-circle of zero longitude to lie in B . Then B contains all points with longitude 0 or π . Because latitude is independent of longitude, it is natural to take the density of θ conditional on B to be $\frac{1}{2}\cos\theta$, the same as the unconditional density.
2. Choose coordinates so that B is the equator (the circle $\theta = 0$). It is natural to take the density of ψ conditional on B to be uniform, the same as the unconditional density.

The two answers are different. Indeed, 1 gives a continuum of different conditional distributions on the circle B , depending on where the poles are fixed, but none of these distributions is uniform on the circle, as 2 is. We will see below that any of these conditional distributions (or any other distribution) can be justified by further information about how the observation was made. It follows that, unless further information is provided, no precise conditional distribution $\underline{P}(\cdot|B)$ is determined by the unconditional distribution \underline{P} .⁴ Your beliefs about ω , apart from knowing that it lies on

the great circle B , should be completely indeterminate. The indeterminacy arises, of course, because $\underline{P}(B) = 0$.

6.10.2 Limits of conditioning events

Borel's paradox shows that it is reasonable for the natural extension $\underline{E}(\cdot|B)$ to be vacuous when $\underline{P}(B) = 0$. To determine non-vacuous conditional probabilities, You must provide further information. One way to do this is to approximate B by events with positive lower probability. If B is one of a continuum of possible observations, it is usually an idealization to suppose that B can be observed precisely.⁵ What is actually observed is some event $B(\delta)$, with non-zero imprecision δ , that contains B . For example, instead of precisely measuring the longitude ψ of ω , You may be able to measure ψ only to within δ .

If the measurement is highly precise (δ is small), it may be convenient to use an idealized model that assumes perfect precision. In this case, the idealized observation B can be regarded as a limit of its neighbourhoods $B(\delta)$ as their imprecision $\delta \rightarrow 0$. Provided $\underline{P}(B(\delta)) > 0$, $\underline{P}(\cdot|B(\delta))$ is determined by \underline{P} through the GBR, and it may be possible to construct $\underline{P}(\cdot|B)$ as the limit (or \liminf) of $\underline{P}(\cdot|B(\delta))$ as $\delta \rightarrow 0$. In this way, further information about the kind of idealization can be used to obtain non-vacuous updated probabilities.

For example, the two conditional distributions in Borel's paradox are justified when B idealizes two different kinds of imprecise measurement.

1. If the longitude ψ is imprecisely measured, to be within δ of 0 or π , then $B(\delta)$ is a band containing B whose width is proportional to $\delta \cos\theta$. (The band narrows to a point at the poles.) Hence the distribution of θ conditional on $B(\delta)$, which is determined by Bayes' rule, has density proportional to $\cos\theta$, and so does the limiting distribution $\underline{P}(\cdot|B)$.
2. If the latitude θ is imprecisely measured, as $|\theta| \leq \delta$, then $B(\delta)$ is a band of width proportional to δ around the equator. The distribution of ψ conditional on $B(\delta)$ is uniform, and so is the limiting distribution $\underline{P}(\cdot|B)$.
3. Index the points of B by α , their angular displacement from a fixed point. Any updated density $g(\alpha)$ can be obtained by assuming that B idealizes a band $B(\delta)$ containing B whose width at α is proportional to $\delta g(\alpha)$.

It is clear that B can be modelled as a limit of discrete observations $B(\delta)$ in many ways, and these can lead to quite different distributions $\underline{P}(\cdot|B)$.

To determine $\underline{P}(\cdot|B)$, it is not enough to specify the idealized partition \mathcal{B} that contains B . If \mathcal{B} partitions the surface into circles of fixed longitude (modulo π), it does not follow that $\underline{P}(\cdot|B)$ is determined by model 1. Suppose, for example, that B is a great circle of zero longitude (mod π), the longitude is

reported to be zero if ω lies within 60 kilometres of B , and otherwise the longitude (mod π) of ω is rounded to the nearest degree. Although a measurement of longitude is reported, the correct model $P(\cdot|B)$ is determined by model 2 rather than 1. What matters is the form of the observation $B(\delta)$, not the idealized partition or σ -field, or the variable through which B is described.⁶

However, specification of a partition \mathcal{B} may often suggest the form of $B(\delta)$. If it is assumed that \mathcal{B} is a refinement of a discrete partition $\mathcal{B}(\delta)$ that contains the actual observation, then $B(\delta)$ must be a union of sets in \mathcal{B} . This assumption will often be justified, especially when \mathcal{B} is constructed as an idealization of $\mathcal{B}(\delta)$, but it is not automatic.

For example, if \mathcal{B} partitions the surface into circles of fixed latitude θ , and $B(\delta)$ is a union of such circles, then $P(\cdot|B)$ is determined by model 2. Here $B(\delta)$ represents an imprecise measurement of one variable (θ), whose precision does not depend on the other variable (ψ). This exemplifies an important type of model that will be considered in 6.10.4.

6.10.3 Conditioning in product spaces

Next we outline a general method for conditioning on precise values of a continuous variable x . Suppose that $\omega = (x, y)$ represents the values of two variables x and y , and $\Omega = \mathcal{X} \times \mathcal{Y}$ is a product space. Let \underline{P} be a coherent lower prevision, defined on a sufficiently large subset of $\mathcal{L}(\Omega)$. We wish to construct previsions $\underline{P}(\cdot|x)$ conditional on x . Formally, $\underline{P}(\cdot|x) = \underline{P}(\cdot|B(x))$ where $B(x) = \{x\} \times \mathcal{Y}$.

Suppose further that $\underline{P}(B(x)) = 0$, and $B(x)$ is regarded as an idealization of events $B(x, \delta)$ for which $\underline{P}(B(x, \delta)) > 0$. Assume that $B(x, \delta)$ are nested neighbourhoods of $B(x)$, which decrease to the limit $B(x)$ as their size δ decreases to zero. When x is a real variable, for example, the neighbourhoods might have the form $B(x, \delta) = \{(u, y): |u - x| \leq \delta g(y)\}$. In general, the ‘precision’ of the neighbourhood, measured here by $\delta g(y)$, may depend on y . (The important case of independence is considered in the following sections.)

Let $A = \mathcal{X} \times A_0$ be an event depending only on y . How can $\underline{P}(A|x)$ be constructed? Provided the physical meaning of y does not depend on the value of x ,⁷ $A \cap B(x, \delta)$ is a δ -neighbourhood of $A \cap B(x)$ in the same way that $B(x, \delta)$ is a δ -neighbourhood of $B(x)$. So it is reasonable to try to approximate $\underline{P}(A|x)$ through the values $\underline{P}(A|B(x, \delta))$, which are determined by \underline{P} through the GBR.

This extends to all gambles Z that are functions of y alone, for which we define $\underline{P}(Z|x) = \liminf_{\delta \rightarrow 0} \underline{P}(Z|B(x, \delta))$. (The liminf will often be a limit.) The justification for this definition is simply that $\underline{P}(Z|x)$ will be at least an

6.10 CONDITIONING ON EVENTS OF PROBABILITY ZERO

approximate lower bound for the correct updated prevision $\underline{P}(Z|B(x, \delta))$, provided the imprecision δ is sufficiently small and Z is a function of y alone.⁸ To ensure separate coherence of $\underline{P}(\cdot|x)$, the definition must be extended to functions of x and y by replacing Z by $Z(x, \cdot)$, which is a function of y alone.⁹

Next we show that the standard method of defining conditional density functions, through a continuous version of Bayes’ rule, can be obtained as a special case of the preceding definition. This requires some substantial assumptions. The most important of these is that the precision of the events $B(x, \delta)$ does not depend on y . In the case where x is a real variable, $B(x, \delta)$ might be identified with an interval $(x - \delta, x + \delta)$.

6.10.4 Bayes’ rule for density functions

Suppose that $\Omega = \mathcal{X} \times \mathcal{Y}$, $x \in \mathcal{X}$, and the following assumptions hold.

- (a) \underline{P} is a linear prevision, defined on a class of measurable gambles by $\underline{P}(Z) = \iint Z(u, y) f(u, y) v(du) \xi(dy)$, where v and ξ are countably additive, σ -finite measures and the joint density function f is measurable and satisfies $f(u, y) \geq 0$, $\iint f(u, y) v(du) \xi(dy) = 1$.
- (b) The marginal density at x , $q(x) = \int f(x, y) \xi(dy)$, is positive.
- (c) $B(x, \delta)$ (subsets of \mathcal{X}) are nested neighbourhoods of x , decrease to the limit $\{x\}$ as δ decreases to zero, are measurable, and satisfy $v(B(x, \delta)) > 0$ and $v(B(x, \delta)) \rightarrow 0$ as $\delta \rightarrow 0$. (We identify $B(x, \delta)$ with $B(x, \delta) \times \mathcal{Y}$, so the precision of the observation $B(x, \delta)$ does not depend on y .)
- (d) The functions $f(\cdot, y)$ are continuous at x , in the sense that $|f(u, y) - f(x, y)| \leq \varepsilon(y, \delta)$ whenever $u \in B(x, \delta)$, where $\int \varepsilon(y, \delta) \xi(dy) \rightarrow 0$ as $\delta \rightarrow 0$.

Define the **conditional density** of y given x to be $g(y|x) = f(x, y)/q(x)$. Then the linear conditional prevision $\underline{P}(\cdot|x)$, which is defined as in section 6.10.3 to be the limit of $\underline{P}(\cdot|B(x, \delta))$ as $\delta \rightarrow 0$, has density $g(\cdot|x)$. That is, $\underline{P}(Z|x) = \int Z(x, y) g(y|x) \xi(dy)$ for all measurable gambles Z .

Proof. Suppose that Z is a function of y alone. Then

$$\underline{P}(ZB(x, \delta)) = \int Z(y) \left(\int_{B(x, \delta)} f(u, y) v(du) \right) \xi(dy).$$

Hence

$$\begin{aligned} & \left| \underline{P}(ZB(x, \delta)) - \int Z(y) f(x, y) \xi(dy) v(B(x, \delta)) \right| \\ &= \left| \int Z(y) \left(\int_{B(x, \delta)} (f(u, y) - f(x, y)) v(du) \right) \xi(dy) \right| \end{aligned}$$

$$\begin{aligned} &\leq \int |Z(y)| \int_{B(x,\delta)} |f(u,y) - f(x,y)| v(du) \xi(dy) \\ &\leq \sup |Z| \int \varepsilon(y, \delta) \xi(dy) v(B(x, \delta)). \end{aligned}$$

Using (d),

$$P(ZB(x, \delta))/v(B(x, \delta)) \rightarrow \int Z(y) f(x, y) \xi(dy) \quad \text{and} \quad P(B(x, \delta))/v(B(x, \delta)) \rightarrow q(x)$$

as $\delta \rightarrow 0$. By (b) and (c), $P(B(x, \delta)) > 0$. Using Bayes' rule,

$$P(Z|B(x, \delta)) = P(ZB(x, \delta))/P(B(x, \delta)) \rightarrow \int Z(y) f(x, y) \xi(dy)/q(x),$$

as required. This defines a linear prevision $P(\cdot|x)$, using linearity of the integral and $\int g(y|x) \xi(dy) = 1$. ◆

This result justifies the use of Bayes' rule for computing conditional density functions, but only under quite restrictive assumptions. To achieve coherence with P , it is not necessary that $P(\cdot|x)$ be defined through the density $g(\cdot|x)$. (Compare with the discrete version of Bayes' rule, which applies when $P(\{x\}) > 0$ and is necessary for coherence.)

Borel's paradox shows that we cannot apply Bayes' rule in an automatic way to obtain conditional densities, because it can produce contradictory answers when applied to different variables which describe the same conditioning event.¹⁰ Bayes' rule should only be applied to those conditioning variables x which satisfy assumption (c).¹¹ The observation must be identified with an imprecise measurement of x , whose precision does not depend on the values of the other variables. Extra information (about the type of imprecision in the real observation) must be supplied to identify the correct conditioning variable x .

The next result states that, when conditional previsions are defined for every value of x through Bayes' rule, these are coherent with the unconditional prevision.

6.10.5 Coherence of Bayes' rule

Suppose that $\Omega = \mathcal{X} \times \mathcal{Y}$ and P is a linear prevision with joint density f , as in 6.10.4(a). When $x \in \mathcal{X}$ and $q(x) > 0$, define $P(\cdot|x)$ to be the linear prevision with density function $g(\cdot|x)$, where $g(y|x) = f(x, y)/q(x)$. When $x \in \mathcal{X}$ and $q(x) = 0$, let $P(\cdot|x)$ be an arbitrary, separately coherent, linear prevision.¹² Then P and $P(\cdot|\mathcal{X})$ are coherent.

Proof. For any measurable gamble Z , use Fubini's theorem¹³ to show that

$$\begin{aligned} P(Z) &= \iint Z(x, y) f(x, y) v(dx) \xi(dy) = \int \left(\int Z(x, y) f(x, y) \xi(dy) \right) v(dx) \\ &= \int P(Z|x) q(x) v(dx) = \int P(Z|x) \left(\int f(x, y) \xi(dy) \right) v(dx) \\ &= \iint P(Z|x) f(x, y) v(dx) \xi(dy) = P(P(Z|\mathcal{X})). \end{aligned}$$

This establishes axiom C15. By 6.5.7, coherence holds. ◆

6.10.6 Lower envelopes

If P is the lower envelope of a class of linear previsions P_γ , each with a joint density function f_γ , then linear conditional previsions $P_\gamma(\cdot|\mathcal{X})$ can be defined through Bayes' rule (as in 6.10.5), and $\underline{P}(\cdot|\mathcal{X})$ can be constructed as the lower envelope of these. Coherence of \underline{P} and $\underline{P}(\cdot|\mathcal{X})$ then follows from 6.10.5 and the lower envelope theorem 7.1.6.

Provided that each P_γ satisfies the assumptions of 6.10.4 and further regularity conditions hold,¹⁴ $\underline{P}(\cdot|x)$ can be justified through the limiting process in 6.10.3, as the limit of $P_\gamma(\cdot|B(x, \delta))$. These results enable us to construct and justify non-vacuous conditional previsions when P is a lower envelope of density functions.¹⁵

6.10.7 Other approaches

There are other ways of constructing non-vacuous conditional previsions when $\underline{P}(B) = 0$. When the upper probability $\bar{P}(B)$ is positive, $\underline{P}(\cdot|B)$ can be defined from P by **regular extension**. This is described in Appendix J. But when $\bar{P}(B) = 0$, as is usual when the partition \mathcal{B} is uncountable, the unconditional prevision P provides absolutely no information about the desirability of gambles contingent on B , as $\underline{P}(BX) = \bar{P}(BX) = 0$ for all gambles X . Then $\underline{P}(\cdot|B)$ cannot be defined in terms of P alone; extra information is needed.

One way of supplying the extra information is to regard B as a limit of discrete events $B(\delta)$, as discussed in this section. Another approach is to model the initial beliefs by some mathematical structure that contains more information than the lower prevision P , especially concerning the desirability of contingent gambles BX . One such model, a class \mathcal{R} of **really desirable** gambles, is described in Appendix F.¹⁶ Conditional previsions can be defined from \mathcal{R} by $\underline{P}(X|B) = \sup\{\mu: B(X - \mu) \in \mathcal{R}\}$. An alternative is to represent previsions by **non-standard real numbers**, and require that all conditioning events B have non-zero (but possibly infinitesimal) probability. Conditional previsions can then be defined as ratios of infinitesimals.¹⁷

Finally, perhaps the simplest way of providing the extra information is to directly assess the conditional previsions $\underline{P}(X|B)$. For example, use the methods of Chapter 4 to assess the unconditional previsions You would adopt if You observed B . It may be easier to do this than to construct a model for initial beliefs that is sufficiently detailed to determine conditional previsions.

6.11 Updating beliefs

To conclude this chapter, we consider the problem of updating beliefs in the light of new information. To apply the GBR, the new information must be modelled as an event B with positive lower probability. Conditional previsions $\underline{P}(\cdot|B)$, which model dispositions to update the initial previsions \underline{P} , are then determined by \underline{P} through the GBR (6.4.1). Does this mean, as some Bayesians have suggested, that Your initial beliefs automatically determine Your beliefs at later times?¹ Other authors have denied that the updated beliefs are constrained in any way by initial beliefs.²

Our conclusions are intermediate between these extremes. The updating strategy suggested by the GBR is to specify initial previsions \underline{P} on a possibility space that is fine enough to model the new evidence as an event B , and to adopt the new unconditional previsions $\underline{P}^B(X) = \underline{P}(X|B)$, which are determined by \underline{P} through the GBR. This strategy can be used quite generally to constrain updated beliefs. There is a role for other updating strategies, not because the updated beliefs constructed through the GBR are unjustified, but because they are often indeterminate.

6.11.1 Difficulties in applying the GBR

The GBR is a consequence of coherence. Essentially, it follows from the updating principle (6.1.6). The following list summarizes the reasons for which the GBR may fail to be applicable or useful as an updating strategy. The first seven problems were discussed earlier (especially in section 6.4.5), and the others are discussed in this section.

1. Conditional previsions need not be constructed from earlier assessments of unconditional previsions. (See sections 6.4.5, 6.5.9 and 6.11.2.)
2. The conditional previsions $\underline{P}(\cdot|B)$ may be highly imprecise, especially when You have done little thinking about gambles contingent on B , perhaps because the occurrence of B was unanticipated. (See the two examples in 6.4, as well as 6.1.3 and 6.11.2.)
3. Further assessments, such as independence, may yield conditional previsions that are more precise than those determined by the GBR (6.1.3, 6.4.5, 6.11.7).

4. You may not have assessed the unconditional previsions needed to apply the GBR (6.4.5, 6.11.5).
5. The GBR is not useful when $\underline{P}(B) = 0$ (6.4.5, 6.8.2, 6.10, Appendix J).
6. You may modify \underline{P} before obtaining any new evidence, by correcting earlier assessments or using new assessment strategies (4.3, 6.1.7).
7. You may realize after observing B that some of Your earlier assessments were unjustified or inadequate (6.1.2).
8. The possible observations B may not partition the initial possibility space Ω (6.4.5, 6.11.3).
9. The new evidence may not be identifiable with a subset B of Ω ; it may tell You more than just that B has occurred (6.4.5, 6.11.4).
10. After representing the new evidence in a refinement of Ω , You must assess various conditional probabilities before the GBR can be applied. Often, the entire model must be constructed after the new evidence is received (6.11.5).
11. The GBR may be more useful when applied to ‘old’ rather than ‘new’ evidence (6.11.6).
12. Other updating strategies may be more useful than the GBR because they yield updated probabilities that are more precise, especially when it is difficult to assess precise probabilities of the new evidence conditional on the states ω (6.11.7). For example, Jeffrey’s rule may be more useful than Bayes’ rule when the new evidence consists of ‘uncertain perceptions’ (6.11.8).

6.11.2 The GBR as an updating strategy

Two steps are involved in updating beliefs by the GBR. The first step is to construct conditional previsions $\underline{P}(\cdot|B)$, through the GBR, from earlier assessments of unconditional previsions. The second step, after B is observed, is to adopt $\underline{P}(\cdot|B)$ as Your new unconditional prevision \underline{P}^B . The earlier theory of coherence suggests that this is a reasonable updating strategy, but there is nothing in the theory that requires You to follow either of these steps.

Concerning the first step, note that the GBR (like the coherence condition from which it follows) is a consistency requirement, not an implicit definition of conditional previsions.³ Like any other consequence of coherence, the GBR can be used as an assessment strategy, when You assess some of the previsions appearing in it and use these to provide information about the others. You might assess $\underline{P}(B(X - \mu))$ for all real μ , and use the GBR to determine $\underline{P}(X|B)$, but it is equally valid to construct \underline{P} from assessments of $\underline{P}(\cdot|B)$ for various events B . What is required is that, however \underline{P} and $\underline{P}(\cdot|B)$ have been constructed, they should satisfy the GBR.

The second point is that the GBR relates \underline{P} to the conditional previsions

$\underline{P}(\cdot|B)$, not to the new previsions \underline{P}^B . Under the ‘updating’ interpretation (6.1.1), $\underline{P}(\cdot|B)$ describes Your initial commitments to adopt new buying prices for gambles after observing B . If B (and nothing else) is observed then these commitments are carried out. Hence \underline{P}^B must dominate $\underline{P}(\cdot|B)$ (6.1.2). But this imposes very weak constraints on \underline{P}^B if $\underline{P}(\cdot|B)$ is highly imprecise, as it may be when determined by the GBR.⁴ Then, after observing B , You may need to do more thinking, and use other assessment strategies, to make \underline{P}^B more precise than $\underline{P}(\cdot|B)$. So there is scope for other updating strategies.

The conditional previsions $\underline{P}(\cdot|B)$ are especially likely to be imprecise when the occurrence of B is unexpected and You have spent little time assessing prior previsions related to B . Suppose that, in a statistical problem, You observe unexpected data (B) which indicate that the parameter lies in the ‘tail’ of Your prior distribution (\underline{P}). Your prior assessments may not determine the tail at all precisely, and the posterior distribution ($\underline{P}(\cdot|B)$) may be highly imprecise.⁵ In that case, You may need to make further assessments, after observing the data, to construct a more precise posterior (\underline{P}^B).⁶ More generally, only after seeing the statistical data can You determine what features of the prior distribution most affect the posterior, and therefore need to be assessed as precisely as possible. So it is reasonable to require only that \underline{P}^B dominates $\underline{P}(\cdot|B)$, and not that they agree.

This does not apply when the unconditional prevision P is linear and $\underline{P}(\cdot|B)$ is determined by Bayes’ rule. Then \underline{P}^B must agree with $\underline{P}(\cdot|B)$. Provided the model P is incorrigible, there is no role for other updating strategies. (These either agree with Bayes’ rule or are inconsistent with P .) That may be why some Bayesians regard Bayes’ rule as the only reasonable way of updating beliefs. The first comment still applies, however. It is just as valid to construct $\underline{P}(\cdot|B)$ in other ways, and use these to construct or modify P .

6.11.3 Possible observations

Next we consider some of the difficulties in applying the GBR, especially in modelling the new evidence as a subset of Ω and assessing its probabilities conditional on the states of interest. First consider item 8 of section 6.11.1. Suppose the new evidence implies that the true state ω belongs to a subset B of Ω . Application of the GBR is justified only if You judge that none of the other observations that could have been made is consistent with B .

Here is a well-known example. Suppose a friend has two children, You are interested in the number of boys, so $\Omega = \{0, 1, 2\}$, and You assess probabilities of the possible values ω . If You learn that at least one child is a boy then You have ‘observed’ $B = \{1, 2\}$, but the information this provides depends on what other observations were possible.⁷ If Your friend answered ‘no’ to the question ‘do you have two girls?’, then the possible

observations partition Ω , the new information can be identified with B , and the GBR can be applied to update Your probabilities. But if he answered ‘no’ to the question ‘is the eldest child a girl?’ then the new information cannot be identified with B . (The answer ‘yes’ is also consistent with B .) In the second case, a version of the GBR can be applied only after further probabilities (of the observation conditional on each state ω) are assessed.⁸

6.11.4 Modelling new evidence as a subset

In the preceding example, learning that the eldest child is a boy tells You that $\omega \in B$, but it also provides extra information that cannot be modelled in terms of Your initial possibility space Ω_0 . To model it, You must refine Ω_0 so that the event that exactly this information is received can be expressed as a subset of the refined space. That is easily done, by defining $\Omega_1 = \{BB, BG, GB, GG\}$, where ‘BG’ means that the eldest child is a boy and the other is a girl.

In most problems, the new evidence You obtain cannot be expressed as a subset of Your initial possibility space.⁹ I have received considerable information in the last year that is relevant to assessing next year’s inflation rate, and this includes much more (such as the forecasts of political leaders) than the information I could have anticipated receiving (such as this year’s inflation figures). I could not have specified in advance a suitable space Ω_0 in which all this information could be expressed. Indeed, it is very difficult to specify the information even after I have received it.¹⁰

6.11.5 A more general updating strategy

If the new evidence cannot be identified with a subset of Your current possibility space Ω_0 , then Your previsions \underline{P} concerning Ω_0 cannot be automatically updated by the GBR. To apply the GBR, You must refine Ω_0 and make further assessments. The first step is to formally represent the new evidence as a point x in an observation space \mathcal{X} , and refine Ω_0 by forming the product space $\Omega_1 = \Omega_0 \times \mathcal{X}$. The new evidence can now be identified with the subset $B = \Omega_0 \times \{x\}$. The next step is to extend \underline{P} , by specifying the previsions of gambles contingent on B . The simplest way to do so is to assess conditional probabilities $\underline{P}(x|\omega)$ and $\bar{\underline{P}}(x|\omega)$ for each ω in Ω_0 . These, together with the Ω_0 -marginal \underline{P} , have a natural extension \underline{E} to $\mathcal{L}(\Omega_1)$. (This was constructed in section 6.7.) The GBR can then be applied to \underline{E} to determine previsions conditional on the new evidence x .

This is a much more general strategy for updating beliefs about Ω_0 , as it can be applied whether or not the new evidence can be identified with a subset of Ω_0 .¹¹ But the required conditional probabilities $\underline{P}(x|\omega)$ and $\bar{\underline{P}}(x|\omega)$,

which are contingent probabilities based only on the old evidence (excluding x), must be assessed after x is observed. The updated previsions are not determined by the assessments made before observing x .

Consider the case in which ω is a statistical parameter.¹² It is common for the sampling models $P(x|\omega)$ to be suggested by the data x , and to be constructed only after seeing the data. Then the entire statistical model, including the parameter space Ω_0 and ‘prior’ prevision \underline{P} on Ω_0 , is constructed after x is observed. Posterior beliefs about ω after observing x are not determined by beliefs prior to the observation.

6.11.6 Old and new evidence

In the statistical example, all the probability assessments are made after the new evidence x is obtained, but based only on the old evidence. It is equally valid to let x denote any other part of the available evidence (not necessarily that obtained most recently), and to assess the probabilities required by the GBR on the basis of the remaining evidence. This further generalizes the assessment strategy outlined in section 6.11.5. It gives a range of different assessment strategies, corresponding to different decompositions of the total evidence into ‘new’ evidence (x) and ‘old’ evidence (everything else). To apply each strategy, You must assess marginal previsions \underline{P} concerning ω , and contingent probabilities $\underline{P}(x|\omega)$ and $\bar{P}(x|\omega)$, all based only on the ‘old’ evidence. The most useful decompositions of the total evidence are those for which these previsions can be assessed most precisely, leading to the greatest precision in updated previsions $\underline{P}(\cdot|x)$.

6.11.7 Comparing strategies

The GBR can be compared with other updating strategies on the same basis, according to the precision of the updated previsions. The general strategy (6.11.5) is likely to be useful when the assessments \underline{P} and $\underline{P}(x|\omega)$ are relatively precise. It will often be useful when the conditional probabilities $P(x|\omega)$ are precise, as when x is a statistical observation whose distribution is determined by ω . In other cases it may be impossible to assess conditional probabilities that are at all precise. That may happen when the evidence x is complicated or difficult to specify, or when it is unclear how x is influenced by the true state ω .¹³ The last case occurs frequently when x represents another person’s opinion about ω .

Suppose, for example, that You have just arrived in a foreign city, You are interested in the event A that it rains tomorrow, Your initial probabilities for A are imprecise because You have little relevant information, and You receive the information x that a weather forecaster assesses precise

probability 0.3 for A . To update Your probabilities for A by the GBR, You must assess $\underline{P}(x|A)$, $\bar{P}(x|A)$, $\underline{P}(x|A^c)$ and $\bar{P}(x|A^c)$. It would be difficult to make precise assessments, especially if You had no previous experience with this forecaster. Even if You could do so, the updated probabilities for A from the GBR would be quite imprecise, because Your initial probabilities are imprecise.¹⁴ However, the alternative updating strategy of simply adopting the expert’s precise probability (or a small interval around it) seems sensible here.¹⁵

This example illustrates that, for some types of new evidence, the GBR is not a useful updating strategy, because it leads to updated previsions that are much less precise than those from alternative strategies.

6.11.8 Jeffrey’s rule¹⁶

Another updating strategy, suggested by Jeffrey (1983), seems more useful than the GBR in problems where the new evidence x is an ‘uncertain perception’. Suppose that You are interested in an event A , You assess precise probabilities $P_0(A|\mathcal{B})$ conditional on a finite partition $\mathcal{B} = \{B_1, \dots, B_n\}$, and You receive new evidence which leads You to update Your \mathcal{B} -marginal probabilities to $P_1(B_j)$ without changing Your conditional probabilities. Then Your updated probability for A is determined by Jeffrey’s rule,

$$P_1(A) = \sum_{j=1}^n P_1(A|B_j)P_1(B_j) = \sum_{j=1}^n P_0(A|B_j)P_1(B_j).$$

Under the assumption that $P_1(A|B_j) = P_0(A|B_j)$, Jeffrey’s rule is a consequence of coherence (6.7.3).

Jeffrey’s rule is useful when Your new evidence cannot be identified with the occurrence of an event, but has the effect of changing Your probabilities for events in the partition \mathcal{B} . Suppose, for example, that You are uncertain about the outcome of an earlier toss of a thumbtack. You think the thumbtack landed pin-up (B_1), but You are not completely certain because You neglected to record the outcome. It may have landed pin-down (B_2). Then Your probability for pin-up on the next toss (A) can be assessed using $P_1(A) = P_0(A|B_1)P_1(B_1) + P_0(A|B_2)P_1(B_2)$, by assessing the probabilities $P_0(A|\mathcal{B})$ of pin-up conditional on the earlier outcome, and Your current probability $P_1(B_1)$ that the earlier outcome was pin-up. To use Bayes’ rule to update $P_0(A)$, You would need to assess probabilities of obtaining Your ‘new evidence’, which consists of Your recollections and uncertainties about the earlier outcome, conditional on the states of interest (such as the occurrence of A , or possible biases of the thumbtack). That seems a very difficult task.

6.11.9 Conclusion

In principle, the general version of the GBR (6.11.5) can always be used to update previsions after receiving new evidence x .¹⁷ It is likely to be a useful updating strategy when the probabilities of obtaining x conditional on the states ω can be assessed relatively precisely. In other cases, when the GBR yields imprecise updated probabilities, there is scope for other updating strategies, such as applying the GBR to different decompositions of the evidence, or using Jeffrey's rule or the general strategy in section 6.7. For some types of evidence, these strategies are more useful than the GBR. Usually, initial beliefs do not determine updated beliefs, but they do impose some constraints on updating, through the GBR.¹⁸

CHAPTER 7

Coherent statistical models

The next two chapters present a theory of statistical reasoning. A statistical observation x , belonging to a sample space \mathcal{X} , is assumed to be generated by one of a class of sampling models $\underline{P}(\cdot|\theta)$, where the unknown parameter θ belongs to a parameter space Θ . You may also assess a prior prevision \underline{P} , representing beliefs about θ or about (θ, x) jointly prior to the observation, and posterior previsions $\underline{P}(\cdot|x)$, representing updated beliefs about θ after observing x .

In this chapter we investigate which statistical models of this sort are coherent. The first task is to define an appropriate concept of coherence, which expresses the rationality constraints that all statistical models should satisfy. The other aims in this chapter are to apply the general concept of coherence to the most important types of statistical problem, to examine the coherence of existing statistical methods, including standard Bayesian inferences, inferences from improper priors and the Neyman–Pearson theory of confidence intervals, and to give examples of statistical models that are coherent but imprecise.

Here is an outline of the chapter. In section 7.1 we define concepts of coherence which are sufficiently general to cover the statistical problems of interest. Six different conditions are defined. Careful attention is given to the differences between them and to their justification as rationality requirements.

The meaning of the parametric sampling model is discussed in section 7.2. Our primary interpretation is that one of the models $\underline{P}(\cdot|\theta)$ is an approximately true description of the process that generates the statistical data. This has several frequentist and propensity versions. The sampling models currently used in statistics are precise, but we also consider imprecise models. The imprecision may be introduced to achieve robustness, to model instability in relative frequencies, or to model physical indeterminacy.

Assuming that a sampling model is specified, different problems of coherence arise according to whether prior or posterior previsions are also assessed. Those cases where no posterior assessments are made are covered by the results of Chapter 6. The basic problem in this chapter, studied in section 7.3, concerns coherence of the sampling model and posterior

previsions. The general coherence conditions in section 7.1 then reduce to simpler axioms S1–S5. These results are applied in section 7.4 to demonstrate the incoherence of various ‘objective’ statistical methods, including Bayesian inferences from improper priors and fiducial inference. These methods are incoherent even in simple problems involving location and scale parameters. For example, if x is an observation from a Normal distribution with unknown mean θ and variance 1, the improper uniform prior generates posterior distributions for θ conditional on x that are Normal with mean x and variance 1. These Normal sampling models and posterior distributions together incur sure loss.

The Neyman–Pearson theory of confidence intervals is discussed in section 7.5. For a confidence-interval estimator with constant coverage probability γ , we examine when it is coherent to adopt γ as a posterior probability of coverage. Again, coherence fails even in simple problems involving location parameters or Normal samples. (The results are related to previous work on ‘relevant subsets’.) The coverage probability γ therefore cannot be interpreted as a posterior probability. Alternative interpretations are considered.

The second type of coherence problem is that in which prior beliefs about θ are assessed, as well as sampling models and posterior previsions. Coherence of these assessments is characterized in section 7.6, through two further axioms S6 and S7. In section 7.7 these results are applied to the case of standard Bayesian models, in which the prior, sampling model and posteriors are all countably additive and have density functions that are related through Bayes’ rule. Such models are always coherent, although they will often be unrealistic because of their precision.

Several methods of constructing coherent inferences from imprecise priors are illustrated in section 7.8. One is to form the lower envelopes of a class of standard Bayesian models. Another is to directly assess an imprecise prior, and define imprecise posterior previsions through the generalized Bayes rule. These results show that, although many widely used statistical methods are not coherent, there is a large class of statistical models which are coherent. Not all of these are lower envelopes of standard Bayesian models.

The third coherence problem, in section 7.9, is that in which prior beliefs are specified concerning θ and x jointly. In this case the prior, sampling model and posterior previsions are coherent just when they are pairwise coherent.

7.1 General concepts of coherence

We first define several generalizations of the coherence concepts in sections 2.3 and 6.3. Suppose that $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_m$ are partitions of a possibility space

Ω , and $\underline{P}(\cdot | \mathcal{B}_i)$ are conditional lower previsions. What does it mean for the m conditional previsions to be coherent?

In statistical problems we can often take $\Omega = \Theta \times \mathcal{X}$ to be a product space, $\mathcal{B}_1 = \{\Omega\}$ to be the trivial partition so that $\underline{P}(\cdot | \mathcal{B}_1)$ is an unconditional lower prevision representing Your beliefs prior to observing data, $\mathcal{B}_2 = \{\{\theta\} \times \mathcal{X}: \theta \in \Theta\}$ so that $\underline{P}(\cdot | \mathcal{B}_2)$ is a sampling model parametrized by θ , and $\mathcal{B}_3 = \{\Theta \times \{x\}: x \in \mathcal{X}\}$ so that $\underline{P}(\cdot | \mathcal{B}_3)$ represents updated beliefs after observing data x . This chapter and the next one will be concerned with statistical models of this form.

In other statistical problems You might assess previsions conditional on several different observables or on several different parameters. For example, You might envisage m successive experiments and take $\underline{P}(\cdot | \mathcal{B}_i)$ to represent Your beliefs after observing the first i experiments. Or You might assess previsions conditional on many different partitions as part of an assessment strategy for constructing an unconditional prevision to represent Your current beliefs. We will therefore define coherence conditions that apply to the general case of previsions conditional on finitely many different partitions.

In deciding what coherence conditions to impose, the central issue must be (as in Chapters 2 and 6): **what conditions can be justified as requirements of rationality?** We start with a requirement of ‘avoiding uniform sure loss’, whose justification follows those given in sections 2.3 and 6.3. We will then see that the justification can be modified slightly to justify stronger conditions.

For simplicity we assume throughout that each domain \mathcal{K}_i of $\underline{P}(\cdot | \mathcal{B}_i)$ is a linear subspace of $\mathcal{L}(\Omega)$.¹ We also assume that each conditional prevision $\underline{P}(\cdot | \mathcal{B}_i)$ is separately coherent; as discussed in section 6.2, this is a minimal coherence requirement. As in section 6.2, write $G(Y|B) = B(Y - \underline{P}(Y|B))$ for the marginal gamble on Y contingent on B , and $G(Y|\mathcal{B}_i) = Y - \underline{P}(Y|\mathcal{B}_i)$ for the two-stage gamble on Y .²

7.1.1 Uniform sure loss

Suppose that $\underline{P}(\cdot | \mathcal{B}_i)$ are separately coherent conditional lower previsions defined on linear spaces \mathcal{K}_i , for $1 \leq i \leq m$. Say that they **avoid uniform sure loss** if

$$\sup_{\omega \in \Omega} \sum_{i=1}^m G(Y_i|\mathcal{B}_i)(\omega) \geq 0 \quad \text{whenever } Y_i \text{ is in } \mathcal{K}_i \text{ for } 1 \leq i \leq m.$$

Justification

Suppose the condition fails. We construct a set of desirable gambles whose sum is a sure loss bounded away from zero. Write $Y = \sum_{i=1}^m G(Y_i|\mathcal{B}_i)$, and

let $Y_i \in \mathcal{K}_i$ be such that $\sup Y < 0$. Let $\delta = -\sup Y/(m+1)$, a positive quantity. By the basic conglomerative principle or its contingent version (section 6.3.3), the gambles $G(Y_i|\mathcal{B}_i) + \delta$ are each desirable. By axiom D3 of section 2.2.3, the finite sum $\sum_{i=1}^m (G(Y_i|\mathcal{B}_i) + \delta) = Y + m\delta$ is desirable. But $Y + m\delta \leq \sup Y + m\delta = -\delta$, which is a sure loss of at least δ . (Hence the term ‘uniform sure loss’.)

This argument can be modified in two ways to justify a stronger condition, by ruling out some losses that are neither ‘uniform’ nor ‘sure’. Regard the desirable gamble $G(Y|\mathcal{B}) + \delta$ as the sum over all conditioning events B in \mathcal{B} of the contingent gambles $G(Y|B) + \delta_B$. We can modify the preceding argument by (i) allowing δ to depend on B , and (ii) taking $\delta = 0$ whenever BY is the zero gamble. The first modification means that the loss we construct will not necessarily be uniform, and the second means that the loss will not necessarily be sure.

Define $S_i(Y)$, called the \mathcal{B}_i -support of Y , to be the subset of \mathcal{B}_i containing all events B such that BY is not identically zero. From our interpretation of contingent previsions, we know that the contingent gamble $G(Y|B) + \delta_B$ is desirable for any positive δ , but not necessarily when $\delta = 0$.³ But if B does not belong to $S_i(Y)$, then BY is the zero gamble and (by separate coherence) $G(Y|B)$ is the zero gamble, which can be regarded as desirable. For that reason, the sets $S_i(Y_i)$ have an important role in the following conditions.

7.1.2 Partial loss

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ are separately coherent conditional previsions defined on linear spaces \mathcal{K}_i , for $1 \leq i \leq m$. Say that they avoid partial loss if, whenever Y_i are in \mathcal{K}_i and are not all zero gambles, there is some event B in $\bigcup_{i=1}^m S_i(Y_i)$ such that

$$\sup_{\omega \in B} \sum_{i=1}^m G(Y_i|\mathcal{B}_i)(\omega) \geq 0.$$

Justification

We modify the justification in section 7.1.1 in the two ways suggested, to construct desirable gambles whose sum is negative whenever any one of them is non-zero. Write $Y = \sum_{i=1}^m G(Y_i|\mathcal{B}_i)$, and suppose $Y_i \in \mathcal{K}_i$ are such that $\sup\{Y(\omega): \omega \in B\} = -\varepsilon(B) < 0$ for each B in $\bigcup_{i=1}^m S_i(Y_i)$. Let $\delta(B) = \varepsilon(B)/(m+1)$. Then $Y \leq 0$, $Y(\omega) < 0$ for all ω in any set of $\bigcup_{i=1}^m S_i(Y_i)$, and $Y(\omega) = 0$ for all other ω . We construct a desirable gamble Z with the same properties, such that $Y \leq Z \leq 0$. (Y itself is not necessarily desirable.)

Let Z_i be the sum of the gambles $B\delta(B)$ over all B in $S_i(Y_i)$. The contingent gamble $G(Y_i|B) + B\delta(B)$ is desirable for each B in $S_i(Y_i)$, and $G(Y_i|B) = 0$ if

7.1 GENERAL CONCEPTS OF COHERENCE

B is not in $S_i(Y_i)$. So the sum of these gambles, $G(Y_i|\mathcal{B}_i) + Z_i$, is desirable by the contingent version of the conglomerative principle. By axiom D3, the gamble $Z = \sum_{i=1}^m (G(Y_i|\mathcal{B}_i) + Z_i) = Y + \sum_{i=1}^m Z_i$ is desirable.

If $\omega \in B$ where $B \in S_i(Y_i)$, then $Z_i(\omega) = \delta(B) = \varepsilon(B)/(m+1) \leq -Y(\omega)/(m+1)$. If $\omega \in B$ and $B \notin S_i(Y_i)$ then $Z_i(\omega) = Y(\omega) = 0$. Thus $Z_i \leq -Y/(m+1)$. Hence $Z = Y + \sum_{i=1}^m Z_i \leq Y - mY/(m+1) = Y/(m+1)$.

Now the gamble Y is negative for all ω in some set of $\bigcup_{i=1}^m S_i(Y_i)$ and Y is zero otherwise, so the same is true of the desirable gamble Z . Although Z is desirable, it cannot possibly be positive and it is negative whenever any of the Y_i is non-zero. That is unreasonable.

This justification might be challenged on three grounds: it uses the contingent version of the conglomerative principle, the constructed loss is not necessarily uniform, and the loss is not necessarily sure. The contingent version of the conglomerative principle was defended in sections 6.3.3 and 6.8.4. Its application here goes beyond that in section 6.3.4, because here it is applied to several partitions \mathcal{B}_i , some of which may not represent observable experiments, and because we allow some of the ‘desirable’ gambles to be the zero gamble.⁴ Nevertheless, this application of the principle does seem to be reasonable.⁵

The existence of a uniform sure loss (7.1.1) is a stronger criticism of specified previsions than is the existence of a partial loss (7.1.2). The question is whether it is irrational to incur losses that are neither sure nor uniform. Uniformity does not seem essential here; a sure loss is a sure loss, and is undesirable whether or not it is bounded away from zero. Nor is it reasonable to accept a gamble, such as Z above, which cannot possibly be positive but can be negative. That seems unreasonable even if the event on which the gamble is negative is in some sense ‘small’, e.g. has zero probability.⁶ Willingness to accept a possible loss of utility, with no possibility of gain, is a type of irrationality.

The different types of loss can be compared with different types of inadmissibility. Consider a choice between two gambles X and Y . (In statistical decision problems, X and Y are the expected utility functions for two decision rules, as in section 3.9.8.) You prefer X to Y just when $Z = X - Y$ is desirable. We say that X is **inadmissible** (Definition 3.9.2) when $Z \leq 0$ and $Z \neq 0$, so that Z is a partial loss. X is **strictly inadmissible** when $Z(\omega) < 0$ for all ω in Ω , so that Z is a sure loss. X is **uniformly inadmissible** when $\sup Z < 0$, so that Z is a uniform sure loss. It is usual in decision making to eliminate all the inadmissible options, whether or not the inadmissibility is strict or uniform. Similarly, we would rule out all inferences that incur partial loss, whether or not the loss is sure or uniform.

We therefore accept the condition of avoiding partial loss as a rationality requirement. We are, however, interested in whether a loss is uniform or

sure, insofar as these are stronger criticisms of specified previsions. The following conditions of avoiding sure loss and avoiding uniform loss are intermediate in strength between avoiding partial loss and avoiding uniform sure loss. It can be verified that previsions incur sure loss just when the desirable gamble Z constructed in justifying 7.1.2 is a sure loss, i.e. $Z(\omega) < 0$ for all ω in Ω . Similarly, they incur uniform loss just when the loss is bounded away from zero on $S(Y)$, the set on which it is non-zero.⁷

7.1.3 Sure or uniform loss

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ are separately coherent conditional previsions defined on linear spaces \mathcal{K}_i . Let $S(Y)$ denote the union of all sets in $\bigcup_{i=1}^m S_i(Y_i)$.

(a) Say that the previsions **avoid sure loss** if, whenever $Y_i \in \mathcal{K}_i$ are such that $S(Y) = \Omega$, there is some event B in $\bigcup_{i=1}^m S_i(Y_i)$ such that

$$\sup_{\omega \in B} \sum_{i=1}^m G(Y_i|\mathcal{B}_i)(\omega) \geq 0.$$

(b) Say that the previsions **avoid uniform loss** if, whenever $Y_i \in \mathcal{K}_i$ are not all zero gambles,

$$\sup_{\omega \in S(Y)} \sum_{i=1}^m G(Y_i|\mathcal{B}_i)(\omega) \geq 0.$$

We have now defined four rationality conditions of different strengths. Examples will be given later in the chapter to show that the differences are important. In particular, many standard statistical methods avoid uniform loss but incur sure loss. (See especially Examples 7.4.4–7.4.6, 7.5.8–7.5.11, 7.6.4.) Simple examples will also be given of inferences that avoid sure loss but incur uniform loss (Examples 7.3.5 and 7.5.12). It is therefore important for the reader to examine critically the justifications for conditions 7.1.1 and 7.1.2. In our view, the argument for avoiding partial loss (the strongest of the four conditions) is compelling.

Next we define weak coherence and coherence as the natural strengthening conditions 7.1.1 and 7.1.2 respectively. These definitions generalize the ideas of coherence in earlier chapters.

7.1.4 Coherence

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ are separately coherent conditional previsions defined on linear spaces \mathcal{K}_i , for $1 \leq i \leq m$.

(a) Say that they are **weakly coherent** if

$$\sup_{\omega \in \Omega} [\sum_{i=1}^m G(Y_i|\mathcal{B}_i) - G(Y_0|B_0)](\omega) \geq 0,$$

7.1 GENERAL CONCEPTS OF COHERENCE

whenever $Y_i \in \mathcal{K}_i$ for all i , and $B_0 \in \mathcal{B}_j$ and $Y_0 \in \mathcal{K}_j$ for some j .

(b) Say that the previsions are **coherent** if, whenever $Y_i \in \mathcal{K}_i$ for all i and $B_0 \in \mathcal{B}_j$ and $Y_0 \in \mathcal{K}_j$ for some j , there is some event B in $\bigcup_{i=1}^m S_i(Y_i) \cup \{B_0\}$ such that

$$\sup_{\omega \in B} [\sum_{i=1}^m G(Y_i|\mathcal{B}_i) - G(Y_0|B_0)](\omega) \geq 0.$$

Justification

Suppose the coherence condition (b) fails. We show that the dispositions represented by $\underline{P}(Y_i|\mathcal{B}_i)$ effectively imply a disposition to pay more than $P(Y_0|B_0)$ for Y_0 contingent on B_0 . Follow the justification of 7.1.2, but define $Y = \sum_{i=1}^m G(Y_i|\mathcal{B}_i) - G(Y_0|B_0)$ and $\delta(B_0) = -\sup\{\bar{Y}(\omega) : \omega \in B_0\}/(m+1)$, so that $\delta(B_0)$ is positive. Define Z_i and Z as in 7.1.2. The argument there shows that Z is desirable and that $Z_i \leq -Y/(m+1)$. Hence

$$Z - G(Y_0|B_0) + B_0\delta(B_0) = Y + \sum_{i=1}^m Z_i + B_0\delta(B_0) \leq 0,$$

using the definition of $\delta(B_0)$. Thus $Z \leq G(Y_0|B_0) - B_0\delta(B_0)$ so the gamble $G(Y_0|B_0) - B_0\delta(B_0)$ is desirable. But this means that You are willing to pay at least $P(Y_0|B_0) + \delta(B_0)$ for Y_0 contingent on B_0 . The specified lower previsions are therefore inconsistent, in the usual sense that one of them can be increased by using the information contained in the others.⁸ This justifies (b). Weak coherence is a consequence of coherence, or it can be justified directly by extending the justification given for 7.1.1.

Coherent previsions avoid partial loss, by taking $Y_0 = 0$ and $B_0 \in \bigcup S_i(Y_i)$ in condition (b). Similarly, weak coherence implies that there is no uniform sure loss. When all the previsions $\underline{P}(\cdot|\mathcal{B}_i)$ are linear we have $G(-Y_0|B_0) = -G(Y_0|B_0)$, from which it follows that coherence is equivalent to avoiding partial loss, and weak coherence is equivalent to avoiding uniform sure loss.⁹ In other cases the coherence conditions are stronger than avoiding loss. Coherence, as defined in 7.1.4(b), is the strongest condition used in this book, although stronger rationality conditions may be justifiable.¹⁰

In the cases of an unconditional prevision or a single partition (studied in Chapters 2 and 6), the new definitions of coherence do reduce to the earlier definitions. It is easily verified that when an unconditional lower prevision \underline{P} is defined on a linear space, coherence of \underline{P} (in the sense of Definition 2.3.3) implies that all of the conditions 7.1.1–7.1.4 are satisfied, with $m = 1$ and $\mathcal{B}_1 = \{\Omega\}$. When a single conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ is specified, separate coherence of $\underline{P}(\cdot|\mathcal{B})$ guarantees that the same conditions are satisfied, with $m = 1$ and $\mathcal{B}_1 = \mathcal{B}$.

Suppose that a coherent unconditional prevision \underline{P} and a separately coherent conditional prevision $\underline{P}(\cdot|\mathcal{B})$ are specified, so $m = 2$. (This is the general case of Chapter 6.) Then conditions 7.1.1–7.1.3 each reduce to the

avoiding sure loss condition in Definition 6.3.2, and each condition in 7.1.4 reduces to the coherence condition in 6.3.2. That is most easily seen from the following theorem, which is useful more generally in verifying coherence.

7.1.5 Reduction theorem

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ ($1 \leq i \leq m$) are defined on linear spaces and separately coherent, and $\mathcal{B}_1 = \{\Omega\}$ is the trivial partition so that $\underline{P} = \underline{P}(\cdot|\mathcal{B}_1)$ is a coherent unconditional prevision. Then the previsions avoid partial loss if and only if

- (a) $\underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ avoid partial loss,
- and (b) $\underline{P}, \underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ avoid uniform sure loss.

(The result remains true if ‘partial’ is replaced by ‘uniform’ or by ‘sure’.) The specified previsions are coherent if and only if

- (a) $\underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are coherent,
- and (b) $\underline{P}, \underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are weakly coherent.

Proof. The two conditions (a) and (b) correspond to the cases where Y_1 is zero or non-zero respectively. When Y_1 is the zero gamble, $S_1(Y_1)$ is empty and $G(Y_1|\mathcal{B}_1) = 0$. Then $\underline{P}(\cdot|\mathcal{B}_1)$ drops out of the definitions, and avoiding partial loss (or coherence, assuming also $j \neq 1$) reduces to condition (a). When Y_1 is non-zero, $S_1(Y_1) = \{\Omega\}$ and then conditions 7.1.2 and 7.1.3 reduce to avoiding uniform sure loss, while coherence reduces to weak coherence. If $j = 1$ in the coherence condition then $B_0 = \Omega$, and coherence again reduces to weak coherence. ◆

As might be expected from earlier results, lower envelopes of coherent specifications are always coherent.

7.1.6 Lower envelope theorem¹¹

Suppose that, for each γ in Γ and $1 \leq i \leq m$, $\underline{P}_\gamma(\cdot|\mathcal{B}_i)$ is a separately coherent, conditional lower prevision defined on a linear space \mathcal{K}_i . Define $\underline{P}(\cdot|\mathcal{B}_i)$ on \mathcal{K}_i to be the lower envelope, $\underline{P}(Y|B) = \inf\{\underline{P}_\gamma(Y|B): \gamma \in \Gamma\}$ when $B \in \mathcal{B}_i$ and $1 \leq i \leq m$. If $\underline{P}_\gamma(\cdot|\mathcal{B}_1), \dots, \underline{P}_\gamma(\cdot|\mathcal{B}_m)$ are coherent for every γ in Γ , then $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are coherent.¹²

Proof. Suppose $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are not coherent. Then there are Y_i, Y_0, B_0 such that $\sup\{Y(\omega): \omega \in B\} = -\varepsilon(B) < 0$ for every B in $\bigcup_{i=1}^m S_i(Y_i) \cup \{B_0\}$, where $Y = \sum_{i=1}^m G(Y_i|\mathcal{B}_i) - G(Y_0|B_0)$. Let $\delta = \varepsilon(B_0)/2$.

Find $\gamma \in \Gamma$ such that $\underline{P}(Y_0|B_0) \geq \underline{P}_\gamma(Y_0|B_0) - \delta$. Write $G_\gamma(Y_0|B_0), G_\gamma(Y_i|\mathcal{B}_i)$ for the marginal gambles corresponding to γ . Then $G(Y_i|\mathcal{B}_i) \geq G_\gamma(Y_i|\mathcal{B}_i)$ and

7.2 SAMPLING MODELS

$$G(Y_0|B_0) \leq G_\gamma(Y_0|B_0) + \delta B_0, \quad \text{hence} \quad Y = \sum_{i=1}^m G_\gamma(Y_i|\mathcal{B}_i) - G_\gamma(Y_0|B_0) \leq Y + \delta B_0.$$

Suppose that $B \in \bigcup_{i=1}^m S_i(Y_i) \cup \{B_0\}$. If $\omega \in B \cap B_0^c$ then $Y_\gamma(\omega) \leq Y(\omega) \leq \sup\{Y(\omega): \omega \in B\} = -\varepsilon(B) < 0$. If $\omega \in B_0$ then $Y_\gamma(\omega) \leq Y(\omega) + \delta \leq \sup\{Y(\omega): \omega \in B_0\} + \delta = -\delta < 0$. Hence $\sup\{Y_\gamma(\omega): \omega \in B\} < 0$. But this contradicts coherence of $\underline{P}_\gamma(\cdot|\mathcal{B}_1), \dots, \underline{P}_\gamma(\cdot|\mathcal{B}_m)$. Thus $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ must be coherent. ◆

This result implies that the lower envelopes of coherent collections of conditional linear previsions form a coherent collection of conditional lower previsions. It is clear from Chapter 6, however, that the converse does not hold: not every coherent collection of conditional lower previsions can be written as a lower envelope of coherent linear collections.¹³

7.2 Sampling models

The rest of this chapter is concerned with parametric statistical problems. These involve models $\underline{P}(\cdot|\theta)$, which represent hypotheses about the process that generated the statistical data. The parameter θ belongs to a set of possible values Θ called the **parameter space**. The family of models $\underline{P}(\cdot|\Theta)$ will be called the **sampling model**. There are no restrictions on the set Θ so that, in principle, this formulation includes ‘nonparametric’ problems, such as those where Θ indexes the family of all probability distributions on the real line.

In this section we examine the philosophical issues arising from the introduction of sampling models. The discussion is relatively brief because our main concern in the present work is with epistemic rather than aleatory probabilities. The basic issue concerns the meaning of the parameter θ and the interpretation of the models $\underline{P}(\cdot|\theta)$. Our primary interpretation of these models is that one of them is an approximately true description of the experiment or phenomenon that generated the data. Several versions of this aleatory interpretation will be described.¹⁴

The sampling models that are currently used in statistics involve precise probabilities. We admit imprecise sampling models, represented by lower previsions $\underline{P}(\cdot|\theta)$. The second issue is the interpretation of imprecise sampling models. Again several interpretations are possible. The imprecision may arise from the inexactness of precise models and the need to make these more robust, or from indeterminacy inherent in the experiment itself.

The final issue is the relationship between aleatory and epistemic probabilities. The coherence conditions in section 7.1 apply to epistemic probabilities, and in the following theory we will interpret the sampling model $\underline{P}(\cdot|\Theta)$ as a model for Your epistemic probabilities contingent on Θ .

The aleatory interpretation of sampling models is used to give meaning to θ and to support the epistemic interpretation, through the principle of direct inference.

7.2.1 Meaningful parameters

For the parameter space Θ to be an admissible possibility space, a unique element θ must be distinguished in an unambiguous way as the ‘true’ value. In that case we will say that the parameter is **meaningful**. It is not required that the true value be observable by some finite procedure. (Indeed, that is why statistical inference is needed.) For example, we regard the chance that a coin lands ‘heads’ on a standard toss as a meaningful parameter, although it cannot be determined exactly through any finite procedure. In this case there is a kind of ‘approximate observability’ or ‘identifiability’, in that the chance of heads can (with high probability) be estimated with arbitrary precision from sufficiently many independent tosses. This kind of identifiability is often useful in clarifying the meaning of a statistical parameter, but it seems enough for that purpose that some method of approximate verification is conceivable. It seems meaningful, for instance, to consider the chance of heads for a coin that has already been destroyed, or the half-life of a radioactive sample that has already fully disintegrated. So meaningful parameters need not be observable and do not require operational definition, although they do require some unambiguous definition.

In some cases θ will have a direct physical interpretation. For example, θ might be a physical quantity such as electrical current that is measured with random error. To simplify interpretation and assessment, it is desirable that parameters (and each component of a vector parameter) should be defined in terms of physical quantities, without reference to sampling probabilities.

However, the chance θ that a coin lands heads is meaningful even though we do not know how to define it in terms of other physical properties of the coin and its tossing method. Most of the following discussion is concerned with sampling models of this type, for which the true value of θ is defined as the index of the true sampling probabilities $P(\cdot|\theta)$.² The parameter θ is meaningful provided these probabilities can be given an aleatory interpretation.

7.2.2 Frequentist interpretations³

There are three rather different versions. The **finite frequency** interpretation takes the probability of an event A , $P(A|\theta)$, to be the relative frequency of

7.2 SAMPLING MODELS

individuals in a real finite population who have the property A .⁴ Here θ simply describes the composition of the population, e.g. the frequencies with which various characteristics occur in it, and so θ does have a direct physical interpretation. It is, in principle, observable. The finite frequency interpretation is natural whenever the statistical experiment under consideration involves sampling from a real population, as in opinion polls, quality control, and other examples of survey sampling.

The **limiting frequency** interpretation of von Mises (1957) takes probability to be the limit of relative frequency in a specific infinite sequence. The infinite sequence must be a ‘collective’: that is, relative frequencies must converge to a limit that is invariant under certain procedures for selecting subsequences. It is somewhat obscure how this interpretation applies to the probabilities involved in sampling models for real, necessarily finite, experiments.

The **hypothetical limiting frequency** interpretation, which seems to be the one adopted by most frequentist statisticians, regards $P(A|\theta)$ as the limit to which the relative frequency of A would converge in a (hypothetical) infinite sequence of repetitions of an experiment.⁵ This refers to a particular experiment rather than a particular sequence of outcomes, and to a conceptual infinite sequence rather than an actual one. The meaning of $P(\cdot|\theta)$ lies in its implications for long-run frequencies in repeated experiments, even if no repetitions are actually performed.

The main objections to this interpretation concern the reference to a ‘long run’ of repeated experiments. In many applications, unlimited repetition, or any repetition at all, may be impossible even in principle. That is true of systems which are not controllable, as in economics or geophysics. Secondly, probability is regarded as a property of a potentially unlimited sequence, and it is unclear how to justify attributing probabilities to single experiments or why they should influence beliefs about single experiments. Finally, the interpretation already refers to a particular experiment, and the further reference to ‘repetitions’ (which are implicitly independent and identical) seems dispensable.

7.2.3 Propensity interpretations⁶

According to these, certain experimental arrangements possess physical dispositions (called **propensities**), which give rise to distributions of aleatory probabilities (called **chances**) in particular experiments (called **trials**). Probability (chance) measures the physical tendency for a particular event, such as ‘heads’, to occur in a particular kind of trial, such as tossing a particular coin in a particular way. Thus chance is an objective property of an experimental arrangement. In statistical problems, the parameter θ is taken to represent a hypothesis about the real chances $P(\cdot|\theta)$ that generated the statistical data.

Chances are not reducible to frequencies, but are related to them through the laws of large numbers.⁷ There is a high chance that relative frequencies in many independent repetitions of an experiment will be close to the chances on each trial. Thus the propensity interpretation can explain the relevance of relative frequencies in repeated experiments, but it is also applicable to experiments that are not repeatable.

Propensities are physical properties and can be linked to other physical properties through scientific laws. In principle, propensities and chances can then be measured through these properties rather than through relative frequencies. For example, a coin that is symmetric in its physical structure must have equal chances for ‘heads’ and ‘tails’.⁸ If the coin is weighted to one side we might be able to predict, from knowledge of physical laws, that this side will have smaller chance than the other of landing face-up, although we know of no physical laws that can be used to predict the chances quantitatively; they can be measured only through relative frequencies of occurrence in repeated tosses.

In other cases propensities are better integrated into scientific laws and theories. Obvious examples are genetics, statistical mechanics and quantum mechanics.⁹ In the theory of radioactive disintegration, the half-life of a radioactive element and the mass of a particular sample determine the chance distribution of the number of disintegrations in any time interval. Half-life is related through theory to nuclear structure and other properties, and this theory can be used to measure half-life by other methods than simply counting frequencies of disintegration.

The example of radioactivity is discussed by Mellor (1971, Ch. 5). He also discusses a more controversial example, of a person’s chance of dying per unit time (called the ‘death risk’). There is a considerable body of theory relating death risk to properties of an individual such as actual age, physiological age, presence of diseases, genetic constitution, blood pressure and environment. This theory can be used to measure a person’s death risk through measurements of the other properties, rather than through observation of the death rate in a population of people who are ‘similar’ to this one.

It seems to us that the propensity interpretation gives a plausible account of the way aleatory probabilities are used in many scientific laws and theories, although some fundamental issues remain unresolved. It does seem essential to a propensity interpretation that propensities be connected to other scientific properties through statistical laws. The aim of statistical analysis is often to discover, test, modify or generalize such laws. However, many applications of statistics are more modest and involve descriptive probability models that are largely specific to the data at hand. In these cases a propensity or frequentist interpretation is less convincing.

There are serious difficulties for a propensity interpretation even in those cases where it seems plausible. The main issues will be discussed in the following subsections.¹⁰

7.2.4 Principle of direct inference

The first issue concerns the connection between aleatory probabilities (chances) and epistemic probabilities. The principle of direct inference asserts that, if Your total evidence about a future trial implies that the chance of event A is $P(A)$, then You should adopt $P(A)$ as Your epistemic probability for A . More generally, if $P(\cdot|\theta)$ is a hypothesis about the chances on the trial, then You should adopt $P(\cdot|\theta)$ as Your epistemic probabilities contingent on θ .¹¹

These are fundamental principles of rationality through which knowledge of, or hypotheses about, physical properties (chances) constrains rational beliefs. We now give two arguments to support the principles.¹²

The first argument relies on our identification of gambles with probability currency (section 2.2.2), which allows epistemic probabilities to be compared directly with chances. Under our behavioural interpretation, the outcome of a gamble is that You win or lose lottery tickets, thereby changing Your chance of winning the prize in a lottery. Consider a gamble $A - \mu$. If You accept this gamble, Your chance of winning the prize will either increase by $\alpha(1 - \mu)$ (if A occurs) or decrease by $\alpha\mu$ (if A does not occur), where α is a positive proportionality constant. Suppose that A has known chance $P(A)$, and occurs independently of the outcome of the lottery. Then accepting the gamble $A - \mu$ increases Your overall chance of winning the prize by $\alpha(1 - \mu)P(A) - \alpha\mu(1 - P(A)) = \alpha(P(A) - \mu)$.

We require the following principle: if accepting a gamble X is known to increase (decrease) Your overall chance of winning the prize, then X is desirable (undesirable). Applying this principle, $A - \mu$ is desirable whenever $\mu < P(A)$ and undesirable whenever $\mu > P(A)$. Hence Your epistemic lower probability $\underline{Q}(A)$, the supremum price for which $A - \mu$ is desirable, must satisfy $\underline{Q}(A) = P(A)$. Similarly, by considering $\mu - A$, $\bar{Q}(A) = P(A)$. We conclude that $\underline{Q}(A)$, Your epistemic probability for A , must be precise and equal to the chance $P(A)$. If the chance $P(A|\theta)$ is unknown, a similar argument, involving the contingent gambles $\{\theta\}(A - \mu)$, establishes that $\underline{Q}(A|\theta) = P(A|\theta)$.

A second, frequentist justification goes as follows. Imagine an infinite sequence of independent repeated trials. Suppose that $P(A)$, the chance of A on each trial, is known. Let \underline{Q} be Your epistemic lower prevision concerning the infinite sequence. Since the trials are identical repetitions, it is reasonable to assume that they are exchangeable under \underline{Q} (Definition 9.5.1),

so that A has the same epistemic probabilities $\underline{Q}(A)$, $\bar{Q}(A)$ on each trial. Let B denote the event that the relative frequency with which A occurs in the sequence converges to $P(A)$. By the law of large numbers, B has chance one. It is therefore natural to assign $\underline{Q}(B) = 1$, expressing certainty that B will occur. But then, by de Finetti's theorem (9.5.4), the trials must be epistemically independent with $\underline{Q}(A) = \bar{Q}(A) = P(A)$. Again, Your epistemic probability for A is precise and agrees with the known chance.

The second argument is less convincing than the first, because of its reference to an infinite sequence of trials. In any case, some version of the principle of direct inference seems necessary to explain how knowledge of chances is relevant to beliefs and decisions. We will use the principle to warrant adoption of sampling models $P(\cdot|\theta)$ as epistemic probabilities, whose coherence with other epistemic probabilities can then be investigated.

7.2.5 Relativized propensities

The next issue is whether chances should be regarded as absolute, intrinsic properties of a unique trial, or as relative to our specification of the trial as one of a particular kind.¹³ Is a person's death risk a property of that person as an individual, or is it a property of the kind of person he is regarded as, e.g. characterized by a set of measurable quantities whose connection to death risk is known?

If propensities are absolute there is a clear conflict with determinism. It is conceivable that whether a person dies in the next year is largely determined by his current physical state, apart from the possibility of accidents or infections, and could be predicted if we had a better understanding of how the body works. Certainly, one would expect increased understanding to lead to different estimates of death risk, perhaps based on new explanatory variables. If death risks are absolute chances, it seems extremely difficult to measure them.

On the other hand, if propensities are relative to the specification of a kind of individual or trial, it is less clear how they can be attributed to single individuals or trials. An individual can be specified in different ways, through different sets of explanatory variables, and these may lead to different death risks.

There are several responses to this dilemma, none of which seems entirely satisfactory. Mellor (1971, Ch. 8) argues that determinism is false simply because we do accept some statistical laws involving propensities as true explanations.¹⁴ The key issue is whether the propensities should be regarded as irreducible. There is a compelling argument for irreducibility in quantum phenomena, but we would be reluctant to claim that statistical laws in other fields are irreducible. In many fields (e.g. geophysics) statistical laws are

regarded as steps toward better models involving further explanatory variables, perhaps leading eventually to deterministic models.¹⁵

A different response is to recognize that we are often more interested in mass behaviour than in individual outcomes. An insurance company, for example, will be interested in predicting death rates in a large population of customers rather than individual deaths. Even if deaths are predictable from detailed measurements made on individuals, that may be of little interest to the insurer.

Such measurements establish initial conditions. Suppose there are known deterministic laws that connect the initial conditions to the outcomes of the trials. These laws explain the differences in outcomes in terms of differences in initial conditions. However, the deterministic laws cannot be used to predict or explain relative frequencies of outcomes in a large sample of trials ('mass behaviour'), unless we can predict or explain the variation in initial conditions. If the initial conditions vary in a random, poorly understood or irregular way in trials of the same kind, e.g. insured individuals or coin tosses, then the deterministic and statistical laws are compatible but useful for different purposes.¹⁶

The difference is particularly striking in the case of earthquakes. The best current models for earthquake occurrence are stochastic. Deterministic models are urgently sought for short-term prediction, but they would not necessarily be useful for modelling long-term frequencies of earthquake occurrence.¹⁷

The relativized interpretation suffices to establish a connection between chances and epistemic probabilities through the principle of direct inference. If we have information only about certain physical characteristics of a person, and also know their death risk relative to these characteristics, then we will use that to determine their chance of dying next year and adopt the chance as our epistemic probability.¹⁸

7.2.6 Approximate sampling models

Few of the sampling models used in statistics are believed to be exactly correct. A realistic theory therefore needs to give some account of the meaning of sampling models that are only approximately true.

To illustrate, suppose we obtain a sample of independent, identically distributed observations. A Normal model $P(\cdot|\theta)$, parametrized by $\theta = (\mu, \sigma^2)$ where μ is the mean and σ^2 is the variance, might be used to analyse the sample, although the true chance distribution that generated the sample is (at best) only approximately Normal. Therefore the 'true' or distinguished value of θ cannot be defined as the index of the true distribution. How else can it be defined?

In some cases θ may have a direct physical interpretation, e.g. μ may be the true value of a physical quantity that is measured with error. In other cases, $\theta = (\mu, \sigma^2)$ can be defined in several ways:

1. Define μ and σ^2 to be the mean and variance of the true distribution P_T , so θ is a partial description of P_T .
2. Define θ to be the limit of estimates¹⁹ in a hypothetical sequence of repetitions of the experiment; e.g. take μ and σ^2 to be the limiting values of the sample mean and sample variance. (With probability one, these agree with the mean and variance of P_T .)
3. Define $P(\cdot|\theta)$ to be the Normal distribution that is closest to P_T under some specified metric.

If the distinguished value of θ is defined in terms of the true distribution P_T , as in (1) and (3), then each θ corresponds to a class \mathcal{M}_θ of possible distributions P_T (e.g. all those with the same mean and variance). Each model in \mathcal{M}_θ is approximated by $P(\cdot|\theta)$. This suggests that approximate sampling models $P(\cdot|\theta)$ can be replaced by imprecise sampling models \mathcal{M}_θ .

As a second example, suppose that n binary observations are modelled as outcomes of (independent, identical) Bernoulli trials, with chance θ . The trials may actually be correlated or have different chances θ_j . In that case, it is natural to define $\theta = n^{-1} \sum_{j=1}^n \theta_j$, the mean chance in the n trials. This value of θ is closest to the true chances under the metric $\sum_{j=1}^n (\theta_j - \theta)^2$, and, with chance one, is the limiting relative frequency of successes in a sequence of repetitions of the experiment. The imprecise sampling model \mathcal{M}_θ contains all chance distributions with mean chance θ .

Most of the sampling models used in practical statistics are only approximate, but the kind of approximation is rarely specified. When it is specified, the approximate (but precise) models $P(\cdot|\theta)$ can be replaced by exact (but imprecise) models \mathcal{M}_θ or $\underline{P}(\cdot|\theta)$, which we consider next.

7.2.7 Robust sampling models

Suppose that, for each θ in Θ , \mathcal{M}_θ is a non-empty, closed, convex class of linear previsions. Then each class \mathcal{M}_θ is uniquely determined by its lower envelope $\underline{P}(\cdot|\theta)$, which is a coherent lower prevision. The **robust sampling models** $\underline{P}(\cdot|\theta)$ are interpreted as lower bounds for the precise underlying chances.²⁰ That is, when θ is the true value, the data are generated by some chance distribution P_T (in \mathcal{M}_θ) that dominates $\underline{P}(\cdot|\theta)$. If θ has a direct physical interpretation then this determines the true model $\underline{P}(\cdot|\theta)$. Otherwise, the true \mathcal{M}_θ is defined as the one that contains P_T ; for θ to be well-defined, it is necessary to assume that different classes \mathcal{M}_θ are disjoint.²¹ So \mathcal{M}_θ are composite statistical hypotheses about the true chances.

Often, the robust sampling model can be described through a **nuisance parameter** ψ , by $\mathcal{M}_\theta = \{P(\cdot|\theta, \psi) : \psi \in \Psi\}$ and $\underline{P}(X|\theta) = \inf\{P(X|\theta, \psi) : \psi \in \Psi\}$. For example, $\theta = (\mu, \sigma^2)$ might be the mean and variance of the true chance distribution, with ψ representing all higher moments. If the true distribution is believed to be approximately Normal, ψ might be restricted to cover only distributions that are close to Normal in shape, so that \mathcal{M}_θ is a ‘neighbourhood’ of the Normal (θ) model.

How is the aleatory model $\underline{P}(\cdot|\Theta)$ related to Your epistemic probabilities contingent on Θ , $\underline{Q}(\cdot|\Theta)$? That depends on Your beliefs about Ψ . Coherence requires that $\underline{Q}(X|\theta)$ agrees with Your lower prevision for $P(X|\theta, \Psi)$ conditional on θ . Hence $\underline{Q}(X|\theta)$ should dominate $\underline{P}(X|\theta)$ in general. Provided You would be completely ignorant about Ψ even if You knew θ , You should adopt the imprecise sampling model $\underline{P}(\cdot|\theta)$ as epistemic probabilities contingent on θ .

The advantages of robust (imprecise) sampling models over approximate (precise) models are that the former may be exactly correct and may have desirable properties of robustness and resistance to outliers. The disadvantages are that they are less tractable and they require more detailed modelling, through specification of the classes \mathcal{M}_θ or the extra parameter ψ .²² However, no probabilities need to be assessed concerning ψ . (The reason for using the imprecise sampling model $\underline{P}(\cdot|\theta)$ rather than the precise model $P(\cdot|\theta, \psi)$ is that there is little or no evidence concerning ψ .) Strict Bayesians, on the other hand, must assess precise probabilities concerning nuisance parameters ψ , and that limits the extent to which they can elaborate or ‘robustify’ their models.²³

The imprecision of a robust sampling model $\underline{P}(\cdot|\theta)$ arises from an absence of knowledge about the precise underlying model $P(\cdot|\theta, \psi)$. This resembles the sensitivity analysis interpretation of epistemic probabilities (section 2.10.4). As with epistemic probabilities, imprecise sampling models $\underline{P}(\cdot|\theta)$ can be given a direct interpretation that does not assume underlying precision. In fact there are versions, for imprecise sampling models, of all the frequentist and propensity interpretations discussed earlier. These are outlined next.

7.2.8 Indeterminate or divergent relative frequencies

First consider a **finite frequency** interpretation of upper and lower probabilities. This is natural whenever individuals in a finite population are classified according to whether they have a property A , but there are some indeterminate cases. Then $\underline{P}(A)$ is the relative frequency of individuals who definitely are A 's, and $\bar{P}(A)$ is the relative frequency of those who possibly are A 's. For example, A might represent an intention to vote for one of two political parties at the next election. There are always indeterminate cases,

who are classified in opinion polls as ‘undecided’. Such indeterminacy is common in surveys asking people for their preference between two options – some may have no preference.

If there are many possible categories, e.g. many political parties, the classification of some individuals may be partially indeterminate, represented by a subset of the categories. This gives rise to the **belief function** models of Dempster and Shafer.²⁴

The **limiting frequency** interpretation takes $\underline{P}(A)$ and $\bar{P}(A)$ to be the lower and upper limits of the relative frequency of A in an actual or hypothetical sequence of repeated trials.²⁵ More generally, the lower and upper previsions of a gamble Y are $\underline{P}(Y) = \liminf_{n \rightarrow \infty} (1/n) \sum_{j=1}^n Y(\omega_j)$ and $\bar{P}(Y) = \limsup_{n \rightarrow \infty} (1/n) \sum_{j=1}^n Y(\omega_j)$, where $\omega_1, \omega_2, \dots$ is the infinite sequence of outcomes. If relative frequencies converge to a limit then the upper and lower previsions agree, and this reduces to the usual limiting frequency interpretation (section 7.2.2). If relative frequencies diverge then \underline{P} is a coherent lower prevision (by section 2.6.8) but is non-linear.²⁶ In statistical problems the limiting upper and lower relative frequencies are unknown, and the sampling models $\underline{P}(\cdot | \theta)$ are interpreted as hypotheses about them.

7.2.9 Imprecise chances

Previous propensity theories have assumed chances to be precise (additive) probabilities, although there is no obvious justification for that assumption.²⁷ As with precise chances, several interpretations of imprecise chances are possible. The most radical interpretation is that imprecise chances represent a kind of ultimate **physical indeterminacy** that is inherent in some phenomena and is not reducible to precise chances. Compare this with the kind of ultimate randomness, represented by precise chances but not reducible to deterministic processes, that appears in quantum mechanics. It seems likely that most physical, social and economic processes are not deterministic, but it is not at all clear that the randomness in any of our current models (apart from quantum theory) is irreducible. Similarly, one might conjecture that most phenomena are ultimately indeterminate, although that would presumably be very difficult to establish for any specific phenomenon. There seems little reason to believe that the world is essentially deterministic, or that it is essentially governed by precise chances.²⁸

A less radical interpretation of imprecise chances might be compatible with underlying precision, just as relativized propensities (section 7.2.5) are compatible with underlying determinism. Trials of the same kind, such as tosses of a coin, may produce different outcomes. The differences in outcome may be explicable, in principle, in terms of differences in initial conditions, but such an ‘explanation’ is worth little unless the variation in initial

7.2 SAMPLING MODELS

conditions is itself identifiable, stable or ‘law-like’. Similarly, imprecision in chances may be due to fluctuations in underlying precise chances which are explicable (in principle) in terms of differences between trials.

Consider, for instance, a long sequence of tosses of a thumbtack. It seems likely that there would be some unavoidable instability in the relative frequencies of outcomes, due to the inability of a human tosser to perfectly reproduce his tossing method. The instability can be ‘explained’ in terms of small variations in tossing method which produce small variations in chances between trials. But such an explanation is useful only if these variations can be identified. It might be impossible in practice to model the detailed variation in chances, and this would impose limits on the precision of a probability model. Imprecise chances would then be used to model the aspects of a process that are stable and identifiable, as opposed to the accidental, uncontrollable or poorly understood variations in precise chances.²⁹ Chance models would be quite precise in the thumbtack example, but might be much less precise for some economic or social processes.

In practice, imprecise models will be required whenever repeated trials produce, or are expected to produce, instability in relative frequencies that cannot be explained in terms of law-like variations in precise chances. Imprecise models do not imply convergence of relative frequencies in independent repetitions, but there remains a connection between chances and relative frequencies. If A is an event in the marginal experiment with upper and lower chances $\bar{P}(A)$, $\underline{P}(A)$, and $r_n(A)$ is its relative frequency in n independent repetitions, then, with lower probability one,

$$\underline{P}(A) \leq \liminf_{n \rightarrow \infty} r_n(A) \leq \limsup_{n \rightarrow \infty} r_n(A) \leq \bar{P}(A).$$

Thus the long-run fluctuations in relative frequencies are bounded by the upper and lower chances.³⁰

7.2.10 Model building

Sampling models need to be chosen or constructed in specific problems, just as epistemic models do. Indeed, formulating a realistic sampling model may be the most important and difficult part of a statistical analysis. Unlike epistemic models, however, sampling models have for a long time been basic tools in statistical practice, and an extensive range of specific models and general model-building techniques has been developed.³¹ The techniques include graphical methods such as scatter plots and probability plots, exploratory methods for smoothing or standardizing data, the use of transformations, robust fitting methods, iterative methods involving analysis of residuals, tests of consistency of models with data, and methods for

estimating density functions. These methods can be used to formulate or modify sampling models after observing the data. Sampling models can often be constructed before seeing the data, based on data obtained in previous studies and on theoretical knowledge or conjectures about underlying processes and mechanisms. Assumptions of physical independence (Chapter 9) usually have an important role in model building.

These standard techniques are normally used to construct precise sampling models. Imprecise models can be constructed by ‘robustifying’ precise ones, as in section 7.2.7. (It might be preferable to construct imprecise sampling models directly from data, but methods for doing so have not yet been developed.) In this chapter we are interested mainly in precise sampling models, since these are well established in statistics, but we also consider the general case of imprecise sampling models.

7.2.11 Other interpretations of sampling models

All the interpretations of sampling models considered so far are aleatory ones, that take the distinguished model to be a description of the process that actually generated the data. Finally we mention four non-aleatory interpretations. The first three are compatible with our theory, insofar as $P(\cdot|\theta)$ can be given an epistemic interpretation.

The **instrumentalist** interpretation regards sampling models as pragmatic tools in modelling that are ‘useful’ rather than ‘true’.³² Their main use is in generating predictive distributions for possible future observations. For example, a life insurance company might use a sampling model involving death risks to predict future claims and to set insurance premiums. An econometric model might be used to forecast future values of economic variables.

One interpretation is that the distinguished parameter θ indexes the model that is ‘most useful’ or ‘most successful’ from the given class Θ , e.g. as measured by its performance in a hypothetical long run of forecasts. Some uses of forecasting and multiple regression techniques, e.g. in econometrics, might be interpreted in this way.

Some objections to an instrumentalist interpretation are that:

1. In practical cases the potential applications of a model are rather vague and open-ended; it may be difficult to define a ‘most successful’ model and to assess prior beliefs about which model would be most successful in hypothetical predictions.
2. Statistics aims at explanation as well as prediction, but instrumentalist models are merely predictive devices.

7.2 SAMPLING MODELS

3. Unless the ‘most successful’ model is regarded as an approximation to a real chance mechanism, its success as a predictive tool appears to be accidental, and there is no reason to expect it to generalize beyond its domain of definition.

The **reductionist** interpretation of de Finetti regards parameters and sampling models as merely convenient mathematical representations for beliefs about a particular set of observables. If, for example, Your beliefs about an infinite sequence of binary observations are exchangeable, they can be represented as a mixture of Bernoulli sampling models. The sampling models have no meaning except through this representation. Thus the meaning of the Bernoulli sampling model is ‘reduced’ to exchangeability of predictive distributions concerning future observations. Some other sampling models can be similarly reduced through concepts of partial exchangeability, which express different kinds of symmetry in beliefs about observables. As a general account of statistical models, the reductionist interpretation seems far-fetched. It will be discussed more fully in section 9.5.

According to the **intersubjective** interpretation of Dawid (1982a), a sampling model describes ‘the area of inter-personal agreement’ amongst the epistemic probability distributions of different persons concerning the same observables.³³ This is an extension of the reductionist interpretation. For example, a group of people who all regarded an infinite sequence of binary observations as exchangeable could accept the Bernoulli sampling model as an intersubjective model, because each person’s beliefs about the observations could be represented as a mixture of Bernoulli models. The objections to the reductionist interpretation (section 9.5.6) apply also to the intersubjective interpretation.

Finally, a purely **descriptive** interpretation of sampling models is common in statistical work. A probability model is regarded as merely an approximate description of observed variation in data, rather than of a random mechanism which might have generated the data. We can describe a sample of real numbers by saying that they approximately follow a Normal distribution with specified mean and variance. There are no implications about underlying chances, about frequencies in a larger population, or about future observations.

If it is decided to fit a specific family of sampling models to the data, a parameter value θ is chosen by the fitting procedure, but this serves merely to summarize the data. The distinguished (fitted) model $P(\cdot|\theta)$ has neither an aleatory nor an epistemic interpretation. An aleatory interpretation is often unrealistic in statistical problems, especially when it is difficult to view the data as outcomes of any kind of experiment, or when there is little understanding of how they were generated.³⁴

7.3 Coherence of sampling model and posterior previsions

In this section we start to examine the implications for statistical problems of the coherence concepts introduced in section 7.1. Here we consider coherence of a parametrized sampling model $\underline{P}(\cdot|\Theta)$ with posterior previsions $\underline{P}(\cdot|\mathcal{X})$. This case is important because of the many attempts to justify ‘objective’ statistical methods which generate precise posterior probabilities without reference to prior beliefs. The attempts include Jeffreys’ theory of improper priors, Fisher’s fiducial theory and (on one interpretation) the Neyman–Pearson theory of confidence intervals. These will be discussed in the two following sections, where we will argue that such attempts are misconceived. The results of this section will be used to show that these methods typically incur sure loss.

First we summarize the notation to be used throughout this and the following chapter.

7.3.1 Statistical notation

The sample space \mathcal{X} is the set of possible outcomes of an experiment or observation. The parameter space Θ indexes a family of hypothetical sampling models $\underline{P}(\cdot|\Theta)$ concerning the experiment. (These have both aleatory and epistemic interpretations, as discussed in section 7.2.) The posterior prevision $\underline{P}(\cdot|\mathcal{X})$ describes beliefs about the true or distinguished parameter value θ after observing experimental outcome x . (This is a purely epistemic interpretation.) Both the sampling model and posterior prevision are conditional previsions; the sampling model is regarded as a contingent prevision, while the posterior is an updated prevision (see section 6.1). To simplify the notation, we will assume throughout that the basic possibility space Ω is the product space $\Theta \times \mathcal{X}$, which consists of all pairs (θ, x) such that $\theta \in \Theta$ and $x \in \mathcal{X}$.¹

The partition of Ω that represents knowledge of θ is $\mathcal{B}_1 = \{A(\theta): \theta \in \Theta\}$ where $A(\theta) = \{(\theta, x): x \in \mathcal{X}\}$. We will simplify the notation of section 7.1 by identifying the set $A(\theta)$ with θ , and the partition \mathcal{B}_1 with Θ . We therefore write $\underline{P}(\cdot|\theta)$ instead of $\underline{P}(\cdot|A(\theta))$, and $\underline{P}(\cdot|\Theta)$ instead of $\underline{P}(\cdot|\mathcal{B}_1)$, for the sampling model parametrized by θ . Similarly we identify the set $B(x) = \{(\theta, x): \theta \in \Theta\}$ with x , identify the partition $\mathcal{B}_2 = \{B(x): x \in \mathcal{X}\}$ with \mathcal{X} , and write $\underline{P}(\cdot|x)$ or $\underline{P}(\cdot|\mathcal{X})$ for the posterior previsions after observing x as the outcome of the experiment. So $\underline{P}(Y|\theta)$ and $\underline{P}(Y|x)$ are real numbers while $\underline{P}(Y|\Theta)$ and $\underline{P}(Y|\mathcal{X})$ are gambles.

In a similar spirit we regard gambles Z on the space Θ as gambles on Ω , by identifying Z with Y such that $Y(\theta, x) = Z(\theta)$ for all x . (Similarly for gambles on the space \mathcal{X} .) All gambles considered can be regarded as gambles on domain Ω .

7.3 COHERENCE OF SAMPLING MODEL AND POSTERIORS

Let $S(Y)$ denote the Θ -support of a gamble Y , defined by $S(Y) = \{\theta \in \Theta: Y(\theta, x) \text{ is non-zero for some } x \in \mathcal{X}\}$. Let $T(Y)$ denote the \mathcal{X} -support of Y , $T(Y) = \{x \in \mathcal{X}: Y(\theta, x) \text{ is non-zero for some } \theta \in \Theta\}$. The sets $S(Y)$ and $T(Y)$ are important here because of their role in the coherence conditions 7.1.2–7.1.4.

Throughout most of this chapter (but not in section 7.5) we will assume that $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ have the same domain \mathcal{F} which consists of all gambles measurable with respect to a product σ -field.²

7.3.2 Assumptions

Let \mathcal{C} be a σ -field of subsets of Θ containing all singletons, and let \mathcal{S} be a σ -field of subsets of \mathcal{X} containing all singletons. Form the product σ -field $\mathcal{C} \times \mathcal{S}$ of subsets of $\Omega = \Theta \times \mathcal{X}$. Let \mathcal{F} denote the linear space of all $\mathcal{C} \times \mathcal{S}$ -measurable gambles. We assume that $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are separately coherent conditional previsions each defined on domain \mathcal{F} . Assume also that, whenever Y is a $\mathcal{C} \times \mathcal{S}$ -measurable gamble, $\underline{P}(Y|\theta)$ is a \mathcal{C} -measurable function of θ and $\underline{P}(Y|x)$ is an \mathcal{S} -measurable function of x , so that both $\underline{P}(Y|\Theta)$ and $\underline{P}(Y|\mathcal{X})$ are in \mathcal{F} .

The following consequences of these assumptions will be used frequently.

- (a) \mathcal{F} is a linear subspace of $\mathcal{L}(\Omega)$ containing all constant gambles.
- (b) If $Y \in \mathcal{F}$ then $S(Y) \in \mathcal{C}$ and $T(Y) \in \mathcal{S}$, so $S(Y)$ and $T(Y)$ can be regarded as gambles in \mathcal{F} .
- (c) If $Y \in \mathcal{F}$ then $\underline{P}(Y|\Theta)$, $\underline{P}(Y|\mathcal{X})$, $G(Y|\Theta)$ and $G(Y|\mathcal{X})$ are all in \mathcal{F} , with $\underline{P}(G(Y|\Theta)|\Theta) = \underline{P}(G(Y|\mathcal{X})|\mathcal{X}) = 0$ by separate coherence, property 6.2.6(j).
- (d) If $Y \in \mathcal{F}$, $\theta \in \Theta$ and $x \in \mathcal{X}$ then $\{\theta\}Y$, $\{x\}Y$, $G(Y|\theta)$ and $G(Y|x)$ are all in \mathcal{F} , with $\underline{P}(G(Y|\theta)|\Theta) = \underline{P}(G(Y|x)|\mathcal{X}) = 0$ by separate coherence.

The rationality conditions introduced in section 7.1 reduce, under the above assumptions, to the following simple axioms.

7.3.3 Avoiding loss theorem

Suppose assumptions 7.3.2 are satisfied. Then $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ avoid partial loss if and only if they satisfy the axiom

- (S1) there is no non-zero Y in \mathcal{F} such that $\underline{P}(Y|\theta) > 0$ for all $\theta \in S(Y)$ and $\bar{P}(Y|x) < 0$ for all $x \in T(Y)$.

The weaker conditions of avoiding sure loss, avoiding uniform loss and avoiding uniform sure loss are equivalent to the following axioms S2, S3, S4 respectively.

- (S2) There is no Y in \mathcal{F} , with either $S(Y) = \Theta$ or $T(Y) = \mathcal{X}$, such that $\underline{P}(Y|\theta) > 0$ for all $\theta \in S(Y)$ and $\bar{P}(Y|x) < 0$ for all $x \in T(Y)$.³
- (S3) There is no non-zero Y in \mathcal{F} such that $\inf\{\underline{P}(Y|\theta): \theta \in S(Y)\} > 0$ and $\sup\{\bar{P}(Y|x): x \in T(Y)\} < 0$.
- (S4) There is no Y in \mathcal{F} such that $\inf\{\underline{P}(Y|\theta): \theta \in \Theta\} > 0$ and $\sup\{\bar{P}(Y|x): x \in \mathcal{X}\} < 0$.

Proof. In terms of the notation of section 7.1 we have $\mathcal{B}_1 = \Theta$, $\mathcal{B}_2 = \mathcal{X}$, $\mathcal{K}_1 = \mathcal{K}_2 = \mathcal{F}$, $S_1(Y) = S(Y)$, $S_2(Y) = T(Y)$. Suppose first that S1 fails for the gamble Y . Let $Z = G(Y|\Theta) + G(-Y|\mathcal{X}) = \bar{P}(Y|\mathcal{X}) - \underline{P}(Y|\Theta)$. When $\theta \in S(Y)$, $Z(\theta, x) \leq -\underline{P}(Y|\theta) < 0$. When $x \in T(Y)$, $Z(\theta, x) \leq \bar{P}(Y|x) < 0$. Thus 7.1.2 fails with $Y_1 = Y$, $Y_2 = -Y$, so $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ incur partial loss.

Conversely, suppose that $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ incur partial loss. By 7.1.2 there are V, W (not both zero) in \mathcal{F} and $Z = G(V|\Theta) + G(W|\mathcal{X})$ such that $\sup Z(\theta, \cdot) < 0$ for all $\theta \in S(V)$ and $\sup Z(\cdot, x) < 0$ for all $x \in T(W)$. Write $C = S(V)$, $D = T(W)$, $U = G(V|\Theta) - G(W|\mathcal{X})$ and $Y = CDU$. By assumptions 7.3.2, each of Z, C, D, U, Y is in \mathcal{F} .

Using $Z \leq 0$ we have $D^c U = D^c G(V|\Theta) = D^c Z \leq 0$. Hence $Y = C(U - D^c U) \geq CU = C(2G(V|\Theta) - Z)$. Using separate coherence, $\underline{P}(Y|\Theta) \geq 2CP(G(V|\Theta)|\Theta) - CP(Z|\Theta) = -CP(Z|\Theta)$. Whenever $\theta \in C$ we have $\underline{P}(Y|\theta) \geq -\bar{P}(Z|\theta) \geq -\sup Z(\theta, \cdot) > 0$. This shows that $S(Y) \supset C$, and $C \supset S(Y)$ by definition of Y , so that $S(Y) = C = S(V)$.

Similarly, $\bar{P}(Y|x) < 0$ whenever $x \in D$, giving $T(Y) = D = T(W)$. Because V and W are not both zero, $S(V) \cup T(W)$ is non-empty, so Y is non-zero. Thus Y violates S1. This establishes that avoiding partial loss is equivalent to S1.

The equivalence of the weaker conditions to S2–S4 can be proved by modifying the preceding proof. When the loss is sure, $S(Y) \cup T(Y) = S(V) \cup T(W) = \Omega = \Theta \times \mathcal{X}$, hence $S(Y) = \Theta$ or $T(Y) = \mathcal{X}$. When the loss is uniform, $\sup\{Z(\theta, x): \theta \in S(V) \text{ or } x \in T(W)\} < 0$, giving $\inf\{\underline{P}(Y|\theta): \theta \in S(Y)\} > 0$ and $\sup\{\bar{P}(Y|x): x \in T(Y)\} < 0$. When the loss is both uniform and sure, take $C = D = \Omega$ and use $\sup Z < 0$. ♦

Axioms S1–S4 are somewhat simpler than the general conditions in section 7.1, and it is easier to understand them as rationality requirements. Suppose first that S1 fails. Then $\underline{P}(Y|\theta) > \frac{1}{2}\bar{P}(Y|\theta) > 0$ for each θ in $S(Y)$. So the gamble $Y - \frac{1}{2}\bar{P}(Y|\Theta)$, in which You pay $\frac{1}{2}\bar{P}(Y|\theta)$ for Y contingent on θ , is desirable. (This is identically zero if θ is not in $S(Y)$.) You will also accept $\frac{1}{2}\bar{P}(Y|\mathcal{X}) - Y$, since You are willing to sell Y for any price greater than $\bar{P}(Y|x)$ after observing x . But the sum of the two gambles is $\frac{1}{2}\bar{P}(Y|\mathcal{X}) - \frac{1}{2}\bar{P}(Y|\Theta)$, which is negative whenever $\theta \in S(Y)$ or $x \in T(Y)$ and is

zero otherwise. So failure of S1 leads You to accept gambles that must lose, overall, if any of them is non-zero.⁴

If S2 fails, $\frac{1}{2}\bar{P}(Y|\mathcal{X}) - \frac{1}{2}\bar{P}(Y|\Theta)$ is negative for all pairs (θ, x) , so that the loss is sure. If S3 fails, the loss is bounded away from zero over the set $S(Y) \cup T(Y)$ on which it is non-zero. If S4 fails, the loss is bounded away from zero over Ω , i.e. is both uniform and sure. While the strongest axiom S1 seems justified as a rationality requirement, violation of S2 or S3 is a more serious failure of rationality, and violation of S4 is still more serious.⁵

The uniform sure loss axiom S4 has the following equivalent forms.⁶

(S4a) There is no Y in \mathcal{F} such that $\inf \underline{P}(Y|\Theta) > \sup \bar{P}(Y|\mathcal{X})$.

(S4b) There is no Y in \mathcal{F} such that $\sup \bar{P}(G(Y|\mathcal{X})|\Theta) < 0$.

The following is a striking example of a uniform sure loss.

7.3.4 Cantelli–Lévy paradox⁷

Let $\Theta = \mathcal{X} = \mathbb{Z}^+$ be the set of positive integers and $\mathcal{F} = \mathcal{L}(\Theta \times \mathcal{X})$. Let Q be a linear prevision which models a ‘uniform distribution’ on \mathbb{Z}^+ (section 2.9.5), so that Q assigns probability zero to each finite set of integers. Define all the conditional linear previsions $P(\cdot|\theta)$ and $P(\cdot|x)$ to agree with Q , i.e. $P(Y|\theta) = Q(Y(\theta, \cdot))$ and $P(Y|x) = Q(Y(\cdot, x))$. De Finetti interprets this as a model for choosing two positive integers θ and x independently and at random, each with distribution Q . Independence means that beliefs about θ are unchanged by knowledge of x , and vice versa. But the model violates S4. To see that, let $A = \{(\theta, x): \theta \leq x\}$ be the event that θ is no greater than x . Then $P(A|\theta) = Q(\{\theta, \theta+1, \dots\}) = 1$ for all $\theta \in \Theta$ because $\{\theta, \theta+1, \dots\}$ has finite complement, whereas $P(A|x) = Q(\{1, 2, \dots, x\}) = 0$ for all $x \in \mathcal{X}$ because $\{1, 2, \dots, x\}$ is finite. Thus S4a fails for the gamble A , and S4 fails for $A - \frac{1}{2}$. The previsions $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ incur uniform sure loss. This is a further indication that de Finetti’s ‘uniform distribution’ on \mathbb{Z}^+ is unreasonable.

Avoiding uniform sure loss is necessary but far from sufficient for reasonable inferences, even when these involve only linear previsions. That is shown by the next example which involves only degenerate probability distributions.

7.3.5 Degenerate example

Let $\Theta = \mathcal{X} = \{1, 2, 3\}$, $\mathcal{F} = \mathcal{L}(\Theta \times \mathcal{X})$. Define $P(\cdot|\theta)$ and $P(\cdot|x)$ to be expectations under degenerate probability distributions, by $P(Y|\theta = j) = Y(j, j)$ for $1 \leq j \leq 3$, $P(Y|x = 1) = Y(2, 1)$, $P(Y|x = 2) = Y(1, 2)$, $P(Y|x = 3) = Y(3, 3)$. That is, $x = \theta$ with probability one conditional on θ . You make the natural inference that $\theta = 3$ (with probability one) when You observe $x = 3$,

but make the unnatural inferences that $\theta = 2$ when You observe $x = 1$, and that $\theta = 1$ when $x = 2$. These probability assessments are (intuitively) highly inconsistent. They do incur uniform loss. To see that they violate S3, let $Y(1, 1) = Y(2, 2) = 1$, $Y(1, 2) = Y(2, 1) = -1$, $Y(\theta, x) = 0$ otherwise. Then $S(Y) = T(Y) = \{1, 2\}$, $P(Y|\theta = 1) = P(Y|\theta = 2) = 1$, $P(Y|x = 1) = P(Y|x = 2) = -1$.

But $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ avoid sure loss. In fact, axiom S2 is automatically satisfied whenever there are $\theta_0 \in \Theta$, $x_0 \in \mathcal{X}$ such that $P(\{x_0\}|\theta_0) = P(\{\theta_0\}|x_0) = 1$, as is the case here with $\theta_0 = x_0 = 3$. This condition is not sufficient for consistency of $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ since it imposes no constraints whatsoever on the other assessments $P(\cdot|\theta)$ and $P(\cdot|x)$. As in this case, the other assessments might be highly inconsistent. This shows that avoiding sure loss (S2) is too weak a requirement.⁸

We next examine coherence of $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. Coherence can be characterized by the following axiom S5, which is a stronger version of S1.⁹

7.3.6 Coherence theorem

Suppose $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are defined on domain \mathcal{F} and satisfy assumptions 7.3.2. They are coherent if and only if they satisfy the two-part axiom:

- (S5) (i) there are no Y in \mathcal{F} and θ_0 in Θ such that $\bar{P}(Y|\theta_0) > 0$,
 $\underline{P}(Y|\theta) > 0$ for all other $\theta \in S(Y)$, and $\bar{P}(Y|x) < 0$ for all $x \in T(Y)$
- (ii) there are no Y in \mathcal{F} and x_0 in \mathcal{X} such that $\underline{P}(Y|x_0) < 0$,
 $\bar{P}(Y|x) < 0$ for all other $x \in T(Y)$, and $\underline{P}(Y|\theta) > 0$ for all $\theta \in S(Y)$.

The proof, which is along the lines of section 7.3.3, is omitted. The two parts of S5 correspond to the choices $B_0 \in \Theta$ or $B_0 \in \mathcal{X}$ in the definition of coherence (7.1.4). The two parts are essentially the same condition but with the roles of θ and x reversed. S5 is somewhat more complicated than S1, but simplifies in some important special cases. Part (i) reduces to S1 when $P(\cdot|\Theta)$ is a linear prevision, and (ii) reduces to S1 when $P(\cdot|\mathcal{X})$ is linear. Thus S5 is equivalent to S1 when both are linear. Also, S5 typically reduces to S1 when both Θ and \mathcal{X} are continuous spaces. (See section 7.6.8 for details.)

As an example of imprecise coherent models, we show next that the vacuous posteriors $\underline{P}(Y|x) = \inf Y(\cdot, x)$ are coherent with any sampling model.

7.3.7 Vacuous posteriors

Suppose $\underline{P}(\cdot|\mathcal{X})$ is vacuous, and $P(\cdot|\Theta)$ is any separately coherent sampling model satisfying assumptions 7.3.2. To establish coherence, verify axiom S5.

7.4 INFERENCES FROM IMPROPER PRIORS

If $\bar{P}(Y|x) = \sup Y(\cdot, x) < 0$ for all $x \in T(Y)$ then $Y \leq 0$, so that $\bar{P}(Y|\theta_0) \leq 0$ for any $\theta_0 \in \Theta$. This establishes S5(i).

For S5(ii), suppose that $\underline{P}(Y|x_0) < 0$ and $\bar{P}(Y|x) < 0$ for all other $x \in T(Y)$. Then $Y(\theta_0, x_0) < 0$ for some $\theta_0 \in \Theta$, $Y(\theta_0, x) < 0$ for all other $x \in T(Y)$, and $Y(\theta_0, x) = 0$ for all $x \in T(Y)^c$. Thus $\theta_0 \in S(Y)$ and $Y(\theta_0, x) \leq 0$ for all $x \in \mathcal{X}$, so that $\underline{P}(Y|\theta_0) \leq 0$. This establishes S5(ii).

Thus it is always coherent, in the absence of non-vacuous prior previsions, to declare complete ignorance about Θ after observing x , irrespective of how x was generated. This result indicates that attempts to generate non-vacuous posterior beliefs from just the sampling model and experimental outcome, without assessments of prior previsions, are likely to prove futile. That issue is examined more closely in the next section.

7.4 Inferences from improper priors

Next we examine the coherence of inferences from improper priors. The results apply also to other statistical methods such as fiducial and structural inference. These methods will be called **objective**, because their inferences depend only on the aleatory sampling model $P(\cdot|\Theta)$, the observation x , and possibly some objective representation of prior ignorance, but not on any subjective prior assessments. It is often emphasized that such methods are appropriate only in cases of prior ignorance concerning Θ . Before showing that the resulting inferences are typically incoherent, we look more closely at a presupposition of these objective approaches, that non-vacuous posterior beliefs are *consistent* with prior ignorance.

7.4.1 Prior ignorance implies posterior ignorance

It is common in statistical problems to feel that initially You have little information about a state θ , so Your prior beliefs about θ are nearly vacuous, but as relevant data accumulate Your beliefs become more precise. On our behavioural interpretation of beliefs, You are initially unwilling to bet concerning θ , but Your upper and lower betting rates tend to converge as the amount of information increases. This phenomenon was illustrated in section 5.3 by the case of Bernoulli trials, with θ the chance of success on each trial. If θ is the chance of an unfamiliar thumbtack landing pin-up, many people would be unwilling to bet that $\theta \geq 0.5$ before observing any tosses, but would become willing to bet after observing 60% pin-ups in a long series of tosses. The appeal of objective statistical methods lies in their apparent ability to model this kind of situation, where prior ignorance is transformed into non-vacuous posterior beliefs.

However, we have already seen (in Example 7.3.7) that the ‘objective’ data (sampling model and observation) alone do not imply any non-vacuous posterior beliefs: they are coherent with a vacuous posterior. The following arguments establish more: the posterior beliefs generated by objective methods are typically in consistent with prior ignorance. The basic idea is that any coherent sampling model $\underline{P}(\cdot|\Theta)$ and posterior previsions $\underline{P}(\cdot|\mathcal{X})$ must be coherent with some prior prevision \underline{P} that describes prior beliefs about Θ . Moreover, there is a minimal such \underline{P} , called the natural extension, which summarizes the implications of $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ for prior beliefs. (These facts will be proved in section 8.2.) The posteriors are consistent with prior ignorance only if the natural extension is vacuous. We give two simple arguments to show that, except in degenerate cases, this can happen only if the posteriors $\underline{P}(\cdot|\mathcal{X})$ are vacuous.

First, suppose there is some non-trivial set A of parameter values whose posterior lower probabilities are bounded away from zero. This will hold whenever the posteriors are non-degenerate. But then, by the conglomerative axiom C8, any prior \underline{P} that is coherent with the posteriors must satisfy $\underline{P}(A) \geq \inf \underline{P}(A|\mathcal{X}) > 0$. You will be prepared to bet on A at some positive rate μ whatever x is observed, so that in effect You are prepared to bet on A at rate μ before the experiment, and Your prior beliefs are effectively non-vacuous.¹

The second argument refers only to a single observation. Suppose there is a non-trivial set A of parameter values and a conceivable observation B , whose sampling probabilities $\underline{P}(B|\theta)$ are bounded away from zero, such that You would be willing to bet on A at some odds after observing B . Here B can be any observation in any conceivable sample space. In the case of Bernoulli trials, B might be the observation that the first ten trials all have the same outcome, and A might be the event that either $\theta \leq 0.25$ or $\theta \geq 0.75$. Most people would be willing to bet on A after observing B , at sufficiently good odds.

The assumptions imply that, for any coherent prior \underline{P} , $\underline{P}(B) \geq \inf \underline{P}(B|\Theta) > 0$ and $\underline{P}(A) \geq \underline{P}(A \cap B) \geq \underline{P}(A|B)\underline{P}(B) > 0$, using properties 4 and 14 of section 6.3.5. In effect, You are initially disposed to bet on A at some positive rate. Again Your prior beliefs about Θ are non-vacuous.²

These arguments show that non-vacuous posterior probabilities effectively imply non-vacuous prior probabilities. Turning this around, prior ignorance implies posterior ignorance.

None of the probabilities in these arguments is assumed to be precise. Indeed, we expect the prior lower probabilities $\underline{P}(A)$ to be often close to zero. They can be equal to zero for some events A , and for other events they may be effectively zero for practical purposes of betting and decision prior to the observation. We have shown only that they cannot be equal

to zero for all A , if You intend to adopt non-vacuous posterior probabilities.³

However, the objective statistical methods considered in this section generate posterior previsions that are not only non-vacuous, but also precise. They will usually imply prior previsions that are precise in at least some respects.⁴ Of course the implied prior beliefs may be reasonable in some problems, but that can be judged only by comparing specific prior previsions with actual prior evidence. This suggests that explicit consideration of prior previsions may be useful, and leads on to the coherence problem in section 7.6.

7.4.2 Improper priors⁵

Let h be a non-negative density function on Θ with respect to a σ -finite measure ξ (often Lebesgue measure). The density is called **improper** when it has infinite integral with respect to ξ , and so cannot be normalized to define a proper prevision. Suppose that linear sampling models $P(\cdot|\theta)$ also have densities $f(\cdot|\theta)$ with respect to a common measure. The role of the improper prior is merely to define a posterior density $g(\cdot|x)$ by formal application of Bayes’ rule. Define $q(x) = \int f(x|\theta)h(\theta)\xi(d\theta)$. Provided $0 < q(x) < \infty$, this generates a proper posterior density $g(\theta|x) = f(x|\theta)h(\theta)/q(x)$, and hence the linear posterior prevision $P(Y|x) = \int Y(\theta, x)g(\theta|x)\xi(d\theta)$.

Statistical models of this type involve a sampling model $P(\cdot|\Theta)$, a posterior prevision $P(\cdot|\mathcal{X})$, and an improper prior density h . Because h is improper it does not determine a coherent prior prevision; it is merely a mathematical device for obtaining a posterior.⁶ But we can still investigate coherence of inferences from the improper prior, by asking whether $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent. The next examples show that standard inferences from improper priors are incoherent. In all the examples, assumptions 7.3.2 hold and all previsions are linear, so that coherence is characterized by Theorem 7.3.3. The first example involves the uniform prior density on the positive integers.

7.4.3 Example of uniform sure loss

Let $\Theta = \mathcal{X}$ be the set of positive integers and $\mathcal{F} = \mathcal{L}(\Theta \times \mathcal{X})$. For each θ in Θ , take the sampling distribution $P(\cdot|\theta)$ to be uniform on the three integers $\{\theta, 2\theta, 2\theta + 1\}$. Consider the uniform prior density (with respect to counting measure) on Θ . This gives posterior probabilities $P(\theta|x) \propto P(x|\theta)$ as a function of θ . If $x > 1$, there are just two distinct integers θ which have $P(x|\theta) > 0$; these are x and the integer part of $x/2$. The posterior distribution $P(\cdot|x)$ therefore assigns probability $\frac{1}{2}$ to each of these integers. For $x = 1$ the posterior distribution is degenerate on $\theta = 1$. Let A be the event that $x = \theta$.

Then $P(A|x=1)=1$, $P(A|x)=\frac{1}{2}$ for $x>1$, but $P(A|\theta)=\frac{1}{3}$ for all θ in Θ . Thus S4 fails for the gamble $Y=0.4-A$, so that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ incur uniform sure loss. The inferences from the uniform prior are incoherent and unreasonable.⁷

7.4.4 Normal distributions

A paradigm of inferences from noninformative priors involves the Normal distribution. Suppose that, under the sampling model $P(\cdot|\theta)$, the observation x has Normal distribution with mean θ and variance 1, written $x \sim N(\theta, 1)$. Here $\Theta = \mathcal{X} = \mathbb{R}$, and \mathcal{F} can be taken to be the class of all Borel-measurable gambles on \mathbb{R}^2 . The standard noninformative prior density is the improper uniform density on the real line, $h(\theta)=1$. This leads to posterior densities $g(\theta|x) \propto f(x|\theta) \propto \exp\{-\frac{1}{2}(x-\theta)^2\}$ as functions of θ , so that the posterior distributions are Normal with mean x and variance 1. This is a particularly simple form of inference: $x \sim N(\theta, 1)$ conditional on θ , and $\theta \sim N(x, 1)$ after observing x .⁸

To see that the inferences are incoherent, consider the event $A = \{(\theta, x): |\theta| \leq |x|\}$, that x is no less than θ in absolute value. Let Φ denote the Normal distribution function. Then, for $\theta \geq 0$, $P(A|\theta) = P(x \geq \theta|\theta) + P(x \leq -\theta|\theta) = \frac{1}{2} + \Phi(-2\theta)$. When $\theta < 0$, $P(A|\theta) = \frac{1}{2} + \Phi(2\theta)$ by a similar argument, so that $P(A|\theta) = \frac{1}{2} + \Phi(-2|\theta|)$ for all real θ . Similarly $P(A|x) = \frac{1}{2} - \Phi(-2|x|)$ for all real x . Thus $P(A|\theta) > \frac{1}{2}$ for all $\theta \in \Theta$ and $P(A|x) < \frac{1}{2}$ for all $x \in \mathcal{X}$. Taking $Y = A - \frac{1}{2}$, this violates axiom S2, so that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ incur sure loss.⁹ Even in this standard example, an improper ‘noninformative’ prior generates incoherent inferences.¹⁰

The sure loss here arises from the desirable gambles $Y - \frac{1}{2}P(Y|\Theta)$ and $\frac{1}{2}P(Y|\mathcal{X}) - Y$, whose sum is $\frac{1}{2}P(Y|\mathcal{X}) - \frac{1}{2}P(Y|\Theta) = \frac{1}{2}P(A|\mathcal{X}) - \frac{1}{2}P(A|\Theta) = -\frac{1}{2}\Phi(-2|x|) - \frac{1}{2}\Phi(-2|\theta|)$. This is not a uniform loss since it tends to zero as both $|x|$ and $|\theta|$ tend to infinity. In fact, we can show that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ do avoid uniform loss.

*Proof.*¹¹ We need to verify S3. Let Y be a Borel-measurable gamble on \mathbb{R}^2 , and write $S = S(Y)$, $T = T(Y)$, $\lambda = \sup|Y|$. Let v denote Lebesgue measure on \mathbb{R} , and let f_0 denote the standard Normal density function so that $g(\theta|x) = f(x|\theta) = f_0(x-\theta)$. Thus

$$P(Y|\theta) = \int_{-\infty}^{\infty} Y(\theta, x) f_0(x-\theta) dx = \int_T Y(\theta, x) f_0(x-\theta) dx$$

and

$$P(Y|x) = \int_{-\infty}^{\infty} Y(\theta, x) f_0(x-\theta) d\theta = \int_S Y(\theta, x) f_0(x-\theta) d\theta.$$

7.4 INFERENCES FROM IMPROPER PRIORS

If $v(T) = 0$ then $P(Y|\theta) = 0$ for all θ so S3 holds. Similarly if $v(S) = 0$. Suppose $v(T) > 0$ and $v(S) > 0$. If $v(S) < \infty$ then $\int_S \int_T Y(\theta, x) |f_0(x-\theta)| dx d\theta \leq \lambda v(S) < \infty$, so Fubini’s theorem can be applied to show that $\int_S P(Y|\theta) d\theta = \int_T P(Y|x) dx$, hence $v(S) \inf\{P(Y|\theta): \theta \in S\} \leq v(T) \sup\{P(Y|x): x \in T\}$, which implies S3.

If $v(S) = \infty$, Fubini’s theorem cannot be applied to change the order of integration, but the following lemma enables us to apply Fubini to integrals over finite intervals which approximate $\int_S P(Y|\theta) d\theta$.

Lemma. Define $\zeta(c) = \int_{-c}^c P(Y|\theta) d\theta - \int_{-c}^c P(Y|x) dx$. There is a constant K such that $|\zeta(c)| \leq K$ for all $c > 0$.

Proof. $\zeta(c) = \int_{-c}^c \int_{-\infty}^{\infty} Y(\theta, x) f_0(x-\theta) dx d\theta - \int_{-c}^c \int_{-\infty}^{\infty} Y(\theta, x) f_0(x-\theta) d\theta dx$. Use Fubini’s theorem to change the order of integration in the second integral, let $z = x - \theta$, and transform $\theta \rightarrow z$ in both integrals. That gives $\zeta(c) = \int_{-\infty}^{\infty} \eta(c, z) f_0(z) dz$, where $\eta(c, z) = \int_{z-c}^{z+c} Y(x-z, x) dx - \int_{-c}^c Y(x-z, x) dx = \int_z^{z+c} Y(x-z, x) dx - \int_{z-c}^z Y(x-z, x) dx$. Hence $|\eta(c, z)| \leq \lambda |\int_z^{z+c} dx| + \lambda |\int_{z-c}^z dx| = 2\lambda|z|$, and $|\zeta(c)| \leq \int_{-\infty}^{\infty} 2\lambda|z| f_0(z) dz = 2\lambda\sqrt{2/\pi}$. This proves the lemma. ◆

To complete the proof, write $S_c = S \cap [-c, c]$, $T_c = T \cap [-c, c]$. Then $v(S_c) \inf\{P(Y|\theta): \theta \in S\} \leq \int_{S_c} P(Y|\theta) d\theta = \int_{-c}^c P(Y|\theta) d\theta = \int_{-c}^c P(Y|x) dx + \zeta(c) = \int_{T_c} P(Y|x) dx + \zeta(c) \leq v(T_c) \sup\{P(Y|x): x \in T\} + K$. But $v(S_c) \rightarrow v(S) = \infty$ as $c \rightarrow \infty$, so that $\inf\{P(Y|\theta): \theta \in S\} > 0$ implies $\sup\{P(Y|x): x \in T\} > 0$, hence S3 holds. ◆

This result shows that the incoherent inferences from the improper prior cannot be exploited to give a loss that is bounded away from zero over Θ and \mathcal{X} . If you believe that $|\theta|$ is very large, so that $|x|$ is also likely to be very large, you might not be worried about the sure loss from the event A , since it is likely to be small. But it is a sure loss! Whatever the values of θ and x , you will lose. That seems unacceptable.

The incoherence can be worse when the improper prior is used for transformations of the parameter θ .¹² Consider the transformation $\psi(\theta) = \sinh(\theta) = \frac{1}{2}\exp(\theta) - \frac{1}{2}\exp(-\theta)$. (This maps \mathbb{R} onto \mathbb{R} and is strictly increasing.) The uniform prior density for ψ gives posterior densities $g(\psi|x) \propto f(x|\theta(\psi))$. If we then transform to get posterior densities for θ ,

$$\begin{aligned} g(\theta|x) &\propto g(\psi(\theta)|x) \frac{d\psi}{d\theta} \propto \exp(-\frac{1}{2}(\theta-x)^2)(\exp(\theta) + \exp(-\theta)) \\ &\propto \exp(x)\exp(-\frac{1}{2}(\theta-x-1)^2) + \exp(-x)\exp(-\frac{1}{2}(\theta-x+1)^2). \end{aligned}$$

So the posterior distribution $P(\cdot|x)$ for θ is a mixture of two Normal distributions with variance one and means $x+1$ and $x-1$. Under each of

these distributions the event $A = \{(\theta, x): |\theta - x| \leq 1\}$ has probability $\frac{1}{2} - \Phi(-2) = 0.477$, so that $P(A|x) = 0.477$ for all $x \in \mathcal{X}$. But $P(A|\theta) = \Phi(1) - \Phi(-1) = 0.683$ for all $\theta \in \Theta$. Thus $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$, generated by the uniform prior for $\sinh(\theta)$, incur uniform sure loss.

7.4.5 Location families

The preceding results for Normal distributions generalize as follows. Say that the sampling model $P(\cdot|\Theta)$ is a **real location family**, and θ is a **real location parameter**, when $\mathcal{X} = \Theta = \mathbb{R}$ and $P(\cdot|\theta)$ has density function of the form $f(x|\theta) = f_0(x - \theta)$ with respect to Lebesgue measure. The sampling model is then determined by the single probability density function f_0 .

When θ is a real location parameter, the recommended noninformative prior is the uniform prior density. This generates posterior densities $g(\theta|x) = f(x|\theta) = f_0(x - \theta)$ with respect to Lebesgue measure. The previous inferences for Normal distributions are a special case of this. Provided the sampling distributions have finite mean ($\int_{-\infty}^{\infty} |z| f_0(z) dz < \infty$), these inferences for real location families avoid uniform loss.¹³ (The proof for Normal distributions used only this property of f_0 .)

However, the inferences do incur sure loss. The event A used for Normal distributions gives a sure loss provided f_0 has unbounded support and median zero. More generally, let β be any strictly decreasing, bounded, measurable function on \mathbb{R} , e.g. $\beta(x) = (1 + \exp(x))^{-1}$. If f_0 has median zero, define the gamble Y by $Y(\theta, x) = \text{sign}(\theta - x)\beta(\min\{\theta, x\})$. (If f_0 has median μ , simply replace θ by $\theta + \mu$ on the right-hand side, which has the effect of translating f_0 to have median zero.) Then $Y(\theta, x) \geq \text{sign}(\theta - x)\beta(\theta)$, with strict inequality whenever $x < \theta$, so that $P(Y|\theta) > \beta(\theta)P(\text{sign}(\theta - x)|\theta) = 0$ for all real θ since f_0 has median zero. Similarly $Y(\theta, x) \leq \text{sign}(\theta - x)\beta(x)$ with strict inequality whenever $x > \theta$, giving $P(Y|x) < 0$ for all real x . Thus S2 fails, so $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ incur sure loss. Inferences from the uniform prior are therefore incoherent for every real location family.

For example, let the sampling distributions $P(\cdot|\theta)$ be uniform on the interval $(\theta - 1, \theta + 1)$. This is a real location family with uniform density $f_0(z) = \frac{1}{2}$ if $|z| < 1$, $f_0(z) = 0$ otherwise. The posterior distribution of θ is uniform on $(x - 1, x + 1)$. These inferences avoid uniform loss (since the mean exists), but they incur sure loss.

These results can be extended to positive scale parameters θ , for which $\mathcal{X} = \Theta = \mathbb{R}^+$ and $f(x|\theta) = \theta^{-1}f_1(\theta^{-1}x)$ for some probability density function f_1 . The recommended noninformative prior density is then $h(\theta) = \theta^{-1}$, which generates posterior densities $g(\theta|x) = x\theta^{-2}f_1(\theta^{-1}x)$. By transforming to $y(x) = \log(x)$ and $\psi(\theta) = \log(\theta)$, we obtain transformed densities $f(y|\psi) = g(\psi|y) = \exp(y - \psi)f_1(\exp(y - \psi))$, so that the model is

7.4 INFERENCES FROM IMPROPER PRIORS

transformed into a real location family determined by the density $f_0(z) = \exp(z)f_1(\exp(z))$. It follows that inferences about scale parameters from the improper prior density $h(\theta) = \theta^{-1}$ also incur sure loss.

7.4.6 Normal observations with unknown mean and variance

Since improper priors lead to a sure loss even in the simple case of location families, it is to be expected that they will typically do so in more complicated problems. Consider a sample of n independent observations $\underline{x} = (x_1, \dots, x_n)$ from a Normal (μ, σ^2) distribution. Here $\Theta = \{(\mu, \sigma): \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ and $\mathcal{X} = \mathbb{R}^n$. The usual noninformative prior density for (μ, σ) is $h(\mu, \sigma) = \sigma^{-1}$, the product of a uniform density for the location parameter μ and density σ^{-1} for the scale parameter σ . This gives posterior densities $g(\mu, \sigma|\underline{x}) \propto \sigma^{-1}f(\underline{x}|\mu, \sigma) \propto \sigma^{-n-1}\exp\{-n[s^2 + (\bar{x} - \mu)^2]/2\sigma^2\}$, where \bar{x} is the sample mean and $s^2 = \sum_{j=1}^n (x_j - \bar{x})^2/n$ is the sample variance.

Again, these inferences avoid uniform loss but incur sure loss. To construct a sure loss, integrate σ out of the posterior densities to obtain marginal density $g(\mu|\underline{x}) \propto [1 + (\bar{x} - \mu)^2/s^2]^{-n/2}$. The posterior distribution of $t(\mu, \underline{x}) = \sqrt{n-1}(\bar{x} - \mu)/s$ is therefore a Student's t -distribution with $n-1$ degrees of freedom.¹⁴ It is well known that this is also the sampling distribution of $t(\mu, \underline{x})$. If we take C to be an event of the form $|t(\mu, \underline{x})| \leq \tau\sqrt{n-1}$, we have $P(C|\underline{x}) = P(C|\mu, \sigma) = \gamma$ for all $(\mu, \sigma) \in \Theta$ and $\underline{x} \in \mathcal{X}$, where γ is a constant (depending only on τ and n) obtained from the t distributions. Here the event C determines intervals $C(\underline{x}) = [\bar{x} - \tau s, \bar{x} + \tau s]$ for μ . These can be regarded both as Bayesian credible intervals with posterior probability $P(C|\underline{x}) = \gamma$, and as frequentist confidence intervals with confidence coefficient $P(C|\mu, \sigma) = \gamma$. We will see in section 7.5.8 that these two sets of probability assessments incur sure loss.¹⁵ (Essentially, this follows from the existence of 'relevant subsets' for the confidence intervals C .)

7.4.7 Other objective methods

The preceding criticisms of improper priors apply also to other objective statistical methods which generate precise posterior distributions. The following methods all generate the posterior distributions for location families that were criticized in sections 7.4.4 and 7.4.5, and the posterior t -distributions in Example 7.4.6. All these methods therefore incur sure loss.¹⁶

1. The theory of noninformative priors: Jeffreys (1983).
2. Fiducial inference: Fisher (1935, 1956).
3. Structural inference: Fraser (1968).

4. Pivotal inference: Barnard (1980).
5. Invariant Haar densities used as improper priors: Berger (1985a).
6. Proper invariant priors: Heath and Sudderth (1978).

Perhaps the fact that these methods all agree in the simple cases of location and scale parameters, and agree also with the more popular theory of confidence intervals, has distracted attention from their incoherence. (Agreement is rare in the foundations of statistics!) But incoherence does seem a very serious criticism of these methods.

7.4.8 Bernoulli trials

We give two further examples to show that inferences from improper priors can sometimes be coherent, and that they can sometimes incur partial loss but not sure loss. In both cases θ is the chance of success in Bernoulli trials.

In the first case, let x be the number of successes in a fixed number (n) of trials. The sampling distributions are binomial (n, θ) , with $\mathcal{X} = \{0, 1, \dots, n\}$ and $P(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, where we use the notation $\binom{n}{x} = n(n-1)\dots(n-x+1)/x(x-1)\dots 1$. Let $\Theta = [0, 1]$, with \mathcal{F} the set of Borel-measurable gambles. Consider the improper prior density $h(\theta) = \theta^{-1}(1-\theta)^{-1}$ on the open interval $(0, 1)$.¹⁷ The posterior density by Bayes' rule is then proportional to $\theta^{x-1}(1-\theta)^{n-x-1}$. This can be normalized to give a proper posterior density provided $1 \leq x \leq n-1$. If $x=0$ or $x=n$, the posterior density is improper and a proper posterior prevision must be chosen in some other way. When $x=0$, the improper posterior density has infinite mass in any neighbourhood of $\theta=0$, so it is natural to take the posterior to be degenerate at $\theta=0$. When $x=n$ we similarly take the posterior to be degenerate at $\theta=1$. Then $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent.¹⁸

Proof. Let Y be any Borel-measurable gamble. We need to verify S1. We have $P(Y|\theta=0) = Y(0,0) = P(Y|x=0)$. If $0 \in T(Y)$ then either $P(Y|x=0) \geq 0$ or $P(Y|\theta=0) < 0$, and S1 holds in each case. Similarly S1 holds if $n \in T(Y)$. Suppose that neither 0 nor n is in $T(Y)$, so $Y(\theta,0) = Y(\theta,n) = 0$ for all $\theta \in [0,1]$. Then

$$\begin{aligned} \int_0^1 P(Y|\theta) \theta^{-1}(1-\theta)^{-1} d\theta &= \int_0^1 \sum_{x=1}^{n-1} Y(\theta,x) \binom{n}{x} \theta^{x-1}(1-\theta)^{n-x-1} d\theta \\ &= \sum_{x=1}^{n-1} \binom{n}{x} \int_0^1 Y(\theta,x) \theta^{x-1}(1-\theta)^{n-x-1} d\theta \\ &= \sum_{x=1}^{n-1} \binom{n}{x} P(Y|x) \int_0^1 \theta^{x-1}(1-\theta)^{n-x-1} d\theta. \end{aligned}$$

7.4 INFERENCES FROM IMPROPER PRIORS

Failure of S1 would force the first expression to be non-negative and the last expression to be negative. So S1 must hold. ◆

Thus the improper prior leads to coherent inferences, provided we take the posteriors to be degenerate when either no successes or no failures are observed. In fact, these degenerate posteriors are essentially the only countably additive posteriors for which $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent. But because they are coherent they have a natural extension to prior beliefs that are coherent with them. In this case they effectively imply that You are initially certain that θ is either zero or one, and that You assign prior probability zero to outcomes $1 \leq x \leq n-1$ for which the improper prior is used to generate the posterior!¹⁹ The posteriors $P(\cdot|\mathcal{X})$ will usually be unreasonable, despite their coherence with $P(\cdot|\Theta)$.

Consider next a different experiment to provide information about θ . Suppose we carry out Bernoulli trials until we observe the first failure. Let x be the number of successes before the first failure, so $\mathcal{X} = \{0, 1, 2, \dots\}$ and the sampling distributions are geometric, $P(x|\theta) = (1-\theta)\theta^x$. Consider first the open parameter space $\Theta = (0, 1)$, with the previous improper prior density $h(\theta) = \theta^{-1}(1-\theta)^{-1}$. The posterior densities for $x \geq 1$ are $g(\theta|x) = x\theta^{x-1}$. If the first trial yields a failure ($x=0$) then the posterior density is improper. But irrespective of how $P(\cdot|x=0)$ is defined, $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ incur sure loss.

To prove that, it is convenient to define $\mu(\theta) = (1-\theta)/\theta$, which is the reciprocal of the mean of the geometric (θ) distribution, so that μ takes values in \mathbb{R}^+ . Let $\beta(\mu) = \frac{1}{2}\mu/(1+\exp(\mu))$. Define the gamble Y in terms of $\mu(\theta)$ by $Y(\theta,x) = (1+\mu)[(1+\mu)^x(\exp(-\mu x) - 2\exp(-2\mu x)) - \beta(\mu)]$ if $x \geq 1$ and $0 < \theta < 1$, and $Y(\theta,x) = 0$ otherwise. It can then be verified that:

- (a) Y is bounded (use the fact that $(1+\mu)\exp(-\mu) < 1$ for all $\mu > 0$),
- (b) $P(Y|\theta) = \beta(\mu) > 0$ for all $0 < \theta < 1$ (by summing geometric series),
- (c) $P(Y|x) = -x \int_0^\infty (1+\mu)^{-x} \beta(\mu) d\mu < 0$ for all $x \geq 1$ (by integrating out the exponential functions).

Because $S(Y) = \Theta$, S2 is violated and there is a sure loss.

The difference from the binomial example is that there \mathcal{X} was finite so that we could change the order of summation over x and integration over θ . In the geometric case, \mathcal{X} is infinite and Fubini's theorem cannot be applied to change the order of summation and integration. The coherence of inferences from the improper prior in the binomial case seems to rely on the finite sample space.

Consider next the geometric experiment with 0 added to the parameter space, so $\Theta = [0, 1]$. If we also take the posterior for $x=0$ to be degenerate at $\theta=0$ there can be no sure loss, by the argument in Example 7.3.5, because

$P(Y|\theta = 0) = Y(0, 0) = P(Y|x = 0)$. But $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are still incoherent, as S1 fails for the gamble Y defined above.²⁰ Because there seems to be no critical difference between the two parameter spaces $(0, 1)$ and $[0, 1]$, this example again suggests that avoiding sure loss (S2) is too weak a requirement and should be strengthened to S1.

Whereas the improper prior density $h(\theta) = \theta^{-1}(1-\theta)^{-1}$ leads to coherent inferences for the binomial experiment, it does not do so for the geometric experiment. But, in practical problems, both experiments may be feasible. It seems unacceptable that the model chosen for prior beliefs should depend on what experiment is later performed, and that suggests that the improper prior is unreasonable even for the binomial experiment.

7.4.9 Relation with proper priors

One way of defending the inferences from improper priors is to show that they are identical to inferences from some proper prior previsions. Consider the example of section 7.4.4, where $P(\cdot|\theta)$ are Normal $(\theta, 1)$ sampling models and $P(\cdot|x)$ are Normal $(x, 1)$ posteriors. The models $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are not coherent, but they can still be extended to a prior prevision \underline{E} , by the type of natural extension described in section 8.2. Their natural extension \underline{E} , defined in 8.2.9, is a translation-invariant lower prevision on the real line. It can be regarded as a coherent replacement for the improper uniform prior.

Alternatively, regard the improper uniform density on \mathbb{R} as a limit of proper uniform densities h_n on the intervals $[-n, n]$.²¹ Each h_n determines a proper prior prevision P_n , and we can construct a lower prevision \underline{P} as the limit $\underline{P}(X) = \liminf_{n \rightarrow \infty} P_n(X)$. This \underline{P} dominates \underline{E} . Again \underline{P} , or one of the translation-invariant linear previsions in $\mathcal{M}(\underline{P})$, can be regarded as a coherent replacement for the improper prior.²²

One criticism of the models \underline{E} and \underline{P} is that they are not fully conglomerable, and none of the linear previsions in $\mathcal{M}(\underline{E})$ or $\mathcal{M}(\underline{P})$ is countably additive. A second criticism is that none of these prior previsions can be coherent with $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$, because these already incur sure loss (section 7.4.4). Rather than supporting the inferences from the improper uniform prior, this seems to discredit the models \underline{E} , \underline{P} , and their dominating linear previsions.

7.4.10 The inappropriateness of improper priors

Our conclusion is that inferences from improper priors are unsound. The idea of generating precise posterior probabilities out of prior ignorance seems misguided (section 7.4.1). The posterior probabilities generated by

improper priors are typically incoherent with the sampling model, even in standard statistical problems involving location parameters or Normal samples (Examples 7.4.4–7.4.6). Many other criticisms of the most popular improper priors were mentioned in section 5.5.

Nor do we see merit in the other objective methods of statistical inference listed in section 7.4.7. Following in the tradition of Bayes and Laplace, they all attempt to conjure precise posterior probabilities out of prior ignorance, a sampling model and an observation; their inferences are indeed as bewildering as conjuring tricks.

What about other objective methods? The Neyman–Pearson theory of confidence intervals is examined in section 7.5. This generates similar inferences to those criticized earlier in this section, although it is less clear how confidence intervals should be interpreted. In our view, the Neyman–Pearson theory does not deal with the real problem of statistical inference: to measure the uncertainty about the unknown parameters of the sampling model after seeing the data.

A more promising alternative is likelihood inference, in which the observed likelihood function from an experiment is regarded as an objective summary of the information provided by the experiment concerning θ . The observed likelihood function does seem to contain all the relevant information, at least when the observation has positive probability (section 8.6). The difficulty is that it is unclear how this information should be used, in the absence of prior probability assessments, to form posterior beliefs and decisions. The argument of sections 7.3.7 and 7.4.1 indicates that the observed likelihood function alone cannot generate non-vacuous posterior beliefs.²³

A third approach, which we favour, is to develop imprecise, but non-vacuous, prior previsions that are reasonable models when there is little prior information. The prior prevision \underline{P} and sampling model $P(\cdot|\Theta)$ then generate posterior previsions $P(\cdot|\mathcal{X})$. The posterior previsions will be imprecise, unlike those generated by other objective methods, and their precision will reflect the amount of information provided by the experiment. The inferences from near-ignorance priors (section 5.3) are one example of this approach.²⁴

7.5 Confidence intervals and relevant subsets

Next we consider how the general concepts of coherence bear upon the Neyman–Pearson theory of confidence intervals.¹ The ideas in this section are related to the literature on relevant subsets, for which key references are Buehler (1959), Pierce (1973) and Robinson (1976, 1979a, 1979b).² It turns out that confidence intervals are typically incoherent under the natural interpretation of their confidence coefficient as a posterior probability. That

is not, in itself, a damning criticism of confidence intervals, as other interpretations are possible.

7.5.1 Basic ideas

Let C denote a zero-one valued function on $\Omega = \Theta \times \mathcal{X}$. For fixed x , $C(x) = \{\theta : C(\theta, x) = 1\}$ is a subset of Θ , and for fixed θ , $\{x : C(\theta, x) = 1\}$ is a subset of \mathcal{X} . Call C a **confidence interval estimator** for θ when $P(C|\theta) = \gamma$ is a constant, not depending on θ .³ The constant γ is called the **confidence coefficient** or **coverage probability**. Typically γ is 0.9 or 0.95.

If x has Normal distribution with mean θ and variance one, for example, then $C = \{(\theta, x) : x > \theta - 1.28\}$ is a 90% confidence interval estimator for θ . In practice we observe the experimental outcome x and then declare that $C(x)$ is a γ confidence interval for θ . After observing $x = 2.35$ (say), we declare that $C(2.35) = (-\infty, 3.63)$ is a 90% confidence interval for θ .

The frequentist interpretation of confidence intervals is that the random set $C(x)$ has aleatory probability γ of covering the true value of θ , whatever that is. Consider a hypothetical sequence of independent observations x_1, \dots, x_n that are generated by a sequence of sampling models $P(\cdot|\theta_1), \dots, P(\cdot|\theta_n)$, and suppose that a γ confidence interval $C(x_j)$ is constructed for each θ_j . As $n \rightarrow \infty$, the relative frequency of those confidence intervals $C(x_j)$ that cover the true value θ_j will converge (with probability one) to γ . Thus, if a particular 90% confidence interval estimator is used in many different problems, it will be ‘successful’ (in covering the true parameter value) in approximately 90% of the problems.

The frequentist interpretation is clear. The problem is how to interpret $C(x)$ after observing x . In the Normal example, what conclusions can we draw from the 90% confidence interval $(-\infty, 3.63)$? Advocates of confidence intervals often emphasize that the confidence coefficient γ is not a posterior probability. But many users of confidence intervals do interpret γ as a reasonable epistemic probability that θ belongs to $C(x)$ after observing x .⁴ They would talk of being ‘90% confident’, or of there being ‘probability 0.9’, that θ is less than 3.63. Before observing x there is probability $P(C|\Theta) = \gamma$ that C will cover θ , so it may seem reasonable to adopt $P(C|x) = \gamma$ as a posterior probability that θ belongs to $C(x)$, whatever x is observed. The question to be answered here is: when are these posterior assessments coherent with the sampling model $P(\cdot|\Theta)$?⁵

It is easy to see that there can be incoherence. Indeed, $C(x)$ can sometimes be the empty set or the whole space Θ , as in the following examples, so that the single probability assessment $P(C|x) = \gamma$ is not even separately coherent.

7.5 CONFIDENCE INTERVALS AND RELEVANT SUBSETS

7.5.2 Normal example

Suppose that x has a Normal $(\theta, 1)$ distribution under $P(\cdot|\theta)$, where Θ is the set of positive reals.⁶ In this case there is an ‘optimal’ one-sided confidence-interval estimator, according to the Neyman–Pearson theory. (This is the ‘uniformly most accurate’ estimator.) The optimal 90% confidence-interval estimator is $C = \{(\theta, x) : x > \theta - 1.28\}$, which yields $C(x) = (0, x + 1.28)$. If the observed value of x is less than -1.28 then the declared confidence interval is empty! In that case You can be certain that $C(x)$ does not cover θ , and it would be absurd to be ‘90% confident’ that $C(x)$ covers θ .

Similarly, if one seeks a lower bound for θ then the optimal one-sided confidence intervals are $C(x) = (\max\{x - 1.28, 0\}, \infty)$. When $x \leq 1.28$, the resulting intervals $C(x)$ cover the whole parameter space. In that case You can be ‘100% confident’ that $C(x)$ covers θ .

The next example is even more extreme.

7.5.3 Uniform example⁷

Suppose that observations x_1, x_2, \dots, x_n are independent and identically distributed, with $n \geq 2$. Under $P(\cdot|\theta)$, x_j takes the value $\theta - 1$ or $\theta + 1$, each with probability $\frac{1}{2}$, where \mathcal{X} and Θ are the set of integers. Let B denote the event that not all the observations x_j are equal. Conditional on B , the estimate $\hat{\theta}(x) = \min\{x_1, \dots, x_n\} + 1$ is equal to θ (with sampling probability one). Define the confidence set $C(x)$ to contain just the point $\hat{\theta}(x)$ when $x \in B$, and to be the empty set when $x \in B^c$. This $C(x)$ covers θ (with probability one) just when B occurs, so its confidence coefficient is $\gamma = P(B|\theta) = 1 - (\frac{1}{2})^{n-1}$. (So $\gamma = 0.5$ when $n = 2$, $\gamma = 0.9375$ when $n = 5$.)

According to the Neyman–Pearson theory, C is the optimal estimator with confidence coefficient γ : it is uniformly most accurate (indeed, the probability of covering false values of θ is uniformly zero), and it has minimum expected size (uniformly in θ). However, whatever x occurs, it is plainly absurd to have confidence γ that $C(x)$ covers the true value of θ . The correct degree of confidence is either one (if B occurs) or zero (if B does not occur).

These examples show that the confidence coefficient γ may not be a sensible measure of posterior ‘confidence’ that $C(x)$ covers θ . Moreover, the fact that ‘optimal’ procedures can produce confidence intervals that are empty, or cover the whole space Θ , suggests that the aim of choosing an estimator with constant coverage probability is misconceived. The inferences that θ belongs to the empty set, or that θ belongs to Θ , seem to be useless, whatever degree of ‘confidence’ is associated with them.

Next we characterize the conditions under which it is coherent to adopt γ as a posterior probability for $C(x)$. The four axioms in the next theorem, which correspond to the four concepts of coherence in section 7.1, involve gambles of the form $Y(C - \gamma)$ where Y is a function of x . These can be regarded as bets on or against the event C at rate γ , with stake $|Y(x)|$ depending on the observation x . Because $P(Y(C - \gamma)|x) = Y(x)(P(C|x) - \gamma) = 0$, the gambles are marginally desirable after observing x . If we subtract a suitable non-negative penalty $Z(x)$, the gamble $Y(C - \gamma) - Z$ will be undesirable after observing x . The theorem states that there is incoherence just when Y and Z can be chosen so that, under each sampling model $P(\cdot|\theta)$, this gamble is (in some sense) favourable.

7.5.4 Confidence theorem⁸

Suppose that

- (a) countably additive, linear previsions $P(\cdot|\Theta)$ are defined on a space \mathcal{F} of measurable gambles and satisfy assumptions 7.3.2⁹
- (b) C is a confidence-interval estimator for θ with confidence coefficient γ , where $C \in \mathcal{F}$ and $0 < \gamma < 1$
- (c) $C(x)$ is non-trivial for each $x \in \mathcal{X}$ (i.e. $C(x)$ is not the empty set or the whole parameter space)¹⁰
- (d) \mathcal{T} is a class of measurable functions of x , with \mathcal{T} contained in \mathcal{F} , and $P(\cdot|\mathcal{X})$ is defined on the linear space $\{YC + X : Y \in \mathcal{T}, X \in \mathcal{F}\}$ by $P(YC + X|\mathcal{X}) = \gamma Y + X$.¹¹

Say that C avoids (partial, uniform, sure) loss if $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ do so.

- (i) For $Y \in \mathcal{T}$, define $J(Y) = \{\theta \in \Theta : Y(x)(C(\theta, x) - \gamma) < 0 \text{ for some } x \in \mathcal{X}\}$. Then C avoids partial loss if and only if it satisfies
- (I1) there are no non-zero functions $Y \in \mathcal{T}$ and $Z \in \mathcal{F}$ such that Z is a function of x , $Z \geq 0$, $Z(x) > 0$ whenever $Y(x) \neq 0$, and $P(Y(C - \gamma) - Z|\theta) > 0$ whenever $\theta \in J(Y)$.
- (ii) C avoids sure loss if and only if it satisfies
- (I2) there are no $Y \in \mathcal{T}$ and $Z \in \mathcal{F}$ such that Z is a function of x , $Z \geq 0$, $Z(x) > 0$ whenever $Y(x) \neq 0$, and $P(Y(C - \gamma) - Z|\theta) > 0$ whenever $\theta \in \Theta$.
- (iii) Define $\tilde{J}(Y) = \Theta$ if Y takes non-zero values that are arbitrarily close to zero, otherwise $\tilde{J}(Y) = J(Y)$. C avoids uniform loss if and only if it satisfies
- (I3) there is no non-zero $Y \in \mathcal{T}$ such that $\inf\{P(Y(C - \gamma)|\theta) : \theta \in \tilde{J}(Y)\} > 0$.
- (iv) C avoids uniform sure loss if and only if it satisfies
- (I4) there is no $Y \in \mathcal{T}$ such that $\inf\{P(Y(C - \gamma)|\theta) : \theta \in \Theta\} > 0$.

Most of the ensuing discussion is concerned with the ‘sure loss’ axiom

7.5 CONFIDENCE INTERVALS AND RELEVANT SUBSETS

I2. If I2 fails, the sure loss can be constructed from the gambles Y and Z as follows. Since $V(\theta) = \frac{1}{2}P(Y(C - \gamma) - Z|\theta)$ is positive for all θ , the gambles $Y(C - \gamma) - Z - V$ are desirable contingent on θ , hence unconditionally desirable (by 6.3.3). Since $P(Z - Y(C - \gamma)|x) = Z(x)$ is positive when $Y(x)$ is non-zero, and $Z - Y(C - \gamma)$ is non-negative when $Y(x)$ is zero, the gamble $Z - Y(C - \gamma)$ is desirable after observing x . The sum of the two desirable gambles is $-V$, a sure loss.

From a frequentist perspective, consider the following **biased betting procedure**.¹² After x is observed, You are paid $\gamma Y(x) + Z(x)$ in return for the gamble $Y(x)C(x)$. When $Y(x)$ is non-zero, this can be regarded as a bet on or against the event that θ belongs to the confidence interval $C(x)$, with stake $|Y(x)|$ and betting rate $\gamma + Z(x)/Y(x)$. You are willing to accept this bet, by the previous argument. But if I2 fails then the betting procedure is unfavourable to You, in the sense that the prevision of Your gain when the true state is θ is $-P(Y(C - \gamma) - Z|\theta)$, which is negative for all possible θ . Thus You will lose (with probability one) from the betting procedure in a long run of repeated trials, whatever the true parameter value, although the bets are acceptable to You after each observation.¹³

If axiom I4 fails, Z can be taken to be a positive constant, and Your gain from the biased betting procedure has uniformly negative prevision conditional on θ . A gamble Y that satisfies the condition $\inf\{P(Y(C - \gamma)|\theta) : \theta \in \Theta\} > 0$ in I4 is called **super-relevant**.¹⁴ The estimator C avoids uniform sure loss if and only if it has no super-relevant gamble.

7.5.5 Relevant subsets

The ‘sure loss’ condition I2 is related to the existence of relevant subsets. Define $P(C|B; \theta) = P(B \cap C|\theta)/P(B|\theta)$ by Bayes’ rule. Following Buehler (1959), call a subset B of \mathcal{X} a **relevant subset** for C when $P(B|\theta) > 0$ for all θ in Θ , and either $\inf\{P(C|B; \theta) : \theta \in \Theta\} > \gamma$ or $\sup\{P(C|B; \theta) : \theta \in \Theta\} < \gamma$. In the first case B is called a **positively biased** relevant subset, and in the second case it is **negatively biased**. The next lemma shows that C incurs sure loss whenever it has relevant subsets.

7.5.6 Lemma

Under the assumptions of Theorem 7.5.4, if there is a relevant subset B for C (where $B \in \mathcal{T}$) then C incurs sure loss.

Proof. Suppose B is a positively biased relevant subset, with $P(C|B; \theta) \geq \gamma + 2\delta$ for all $\theta \in \Theta$, where $\delta > 0$. Then

$$P(B(C - \gamma) - \delta B|\theta) = P(B|\theta)(P(C|B; \theta) - \gamma - \delta) \geq \delta P(B|\theta) > 0 \quad \text{for all } \theta \in \Theta.$$

Thus I2 fails for $Y = B$ and $Z = \delta B$. If B is negatively biased with $P(C|B;\theta) \leq \gamma - 2\delta$, a similar calculation shows that I2 fails for $Y = -B$ and $Z = \delta B$. ♦

When B is a relevant subset, the sure loss can be constructed from gambles contingent on B . For positively biased B , I can construct a biased betting procedure by betting that $C(x)$ covers θ whenever x is in B , at rate $\gamma + \delta$, and not betting for other x . Provided $\inf\{P(B|\theta): \theta \in \Theta\} > 0$, any relevant subset B is super-relevant and yields a uniform sure loss.

For example, if $C(x) = \Theta$ when $x \in B$ and $P(B|\theta) > 0$ for all $\theta \in \Theta$ then B is a positively biased relevant subset. If $C(x)$ is empty when $x \in B$ and $P(B|\theta) > 0$ for all $\theta \in \Theta$ then B is a negatively biased relevant subset. In Example 7.5.3, B is positively biased, since $P(C|B;\theta) = 1$ for all θ , and B^c is negatively biased, since $P(C|B^c;\theta) = 0$ for all θ . Here both B and B^c are super-relevant subsets.

7.5.7 Frequentist meaning

The existence of relevant subsets is disturbing to frequentists for the following reason. The frequentist interpretation of the confidence coefficient γ is that it is the limiting relative frequency with which the random set $C(x)$ covers the true value of θ , in a sequence of repeated trials governed by θ . But if there is a relevant subset B , so that $P(C|B;\theta)$ is bounded away from γ , this defines a **recognizable subsequence** of trials (those on which B occurs) in which the limiting relative frequency of coverage is bounded away from γ . If the observation x belongs to B then the conditional coverage probabilities $P(C|B;\theta)$ seem more relevant than the unconditional probability $P(C|\theta)$ to inferences concerning θ . This illustrates the fundamental problem for a frequentist theory of statistical inference: what subsequence of trials, or conditioning event B , should be used to evaluate sampling probabilities? In many examples, such as 7.5.3 and 7.5.11, for the hypothetical relative frequencies to be relevant in drawing inferences from the observed data, some kind of conditioning seems essential.¹⁵

The idea of relevant subsets was used by Fisher (1956a) to criticize the Neyman–Pearson theory of confidence intervals.¹⁶ Ironically, relevant subsets also exist for Fisher's own fiducial intervals. The classic example is the following, where fiducial intervals agree with the confidence intervals based on Student's t -distribution.¹⁷

7.5.8 Normal observations with unknown mean and variance

Let \tilde{x} denote a sample of n (≥ 2) independent observations from a Normal (μ, σ^2) distribution, so that $\Theta = \{(\mu, \sigma): \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$, $\mathcal{X} = \mathbb{R}^n$. The standard

confidence intervals for μ are based on the t -distribution of $t(\mu, \tilde{x}) = \sqrt{n-1}(\bar{x} - \mu)/s$ conditional on (μ, σ) , where \bar{x} is the sample mean and $s^2 = \sum_{j=1}^n (x_j - \bar{x})^2/n$ is the sample variance. So C is the event that $|\bar{x} - \mu| \leq ts$ for some specified t , and γ is the probability that $|t| \leq \tau \sqrt{n-1}$ when t has a t -distribution with $n-1$ degrees of freedom. The estimator C has some optimality properties according to the Neyman–Pearson theory.¹⁸

Let $B = \{\tilde{x}: |\bar{x}| \leq \beta s\}$. This is the event that, using the standard t -test, we accept the null hypothesis $\mu = 0$ at a certain significance level.¹⁹ Intuitively, when B occurs, s is more likely to be large, and the confidence interval is more likely to cover μ . Buehler and Fedderson (1963) and Brown (1967) show that B is a positively biased relevant subset for C , provided β is sufficiently large. By Lemma 7.5.6, these standard confidence intervals incur sure loss.²⁰

For the case $n = 2$, Brown (1967) shows that the choice $\beta = \tau + (1 + \tau^2)^{1/2}$ yields conditional coverage probabilities $P(C|B; \mu, \sigma) \geq 2\gamma/(1 + \gamma)$ for all values of μ and σ . For example, the lower bound is $\frac{2}{3}$ when $\gamma = 0.5$, and it is 0.974 when $\gamma = 0.95$. The conditional coverage probabilities are quite different from the unconditional ones.²¹

The standard noninformative prior for (μ, σ) (7.4.6) generates a posterior t -distribution for $t(\mu, \tilde{x})$, and hence yields Bayesian credible intervals for μ that agree with the confidence intervals and have posterior probability γ . It follows that these improper Bayesian inferences also incur sure loss. The intervals $C(x)$ are very often used in practice by both frequentists and Bayesians, so this is a case where ideas of coherence can be used to criticize standard practice.²²

7.5.9 Location families

Consider the real location families defined in section 7.4.5. Assume that the sampling densities $f(x|\theta) = f_0(x - \theta)$ have finite mean. Let A be any measurable set of reals such that $\int_A f_0(y) dy = \gamma$. Define $C = \{(\theta, x): x - \theta \in A\}$. Then $P(C|\theta) = \int_{x-\theta \in A} f(x|\theta) dx = \int_A f_0(y) dy = \gamma$, so that C is a confidence interval estimator for θ with confidence coefficient γ .

If You adopt a uniform prior density for θ , giving posterior densities $g(\theta|x) = f_0(x - \theta)$, we know from section 7.4.5 that the posterior previsions and sampling models avoid uniform loss. Under these posteriors $P(C|x) = \int_{x-\theta \in A} g(\theta|x) d\theta = \int_A f_0(y) dy = \gamma$, so that $C(x)$ does have posterior probability γ . For location families the confidence intervals $C(x)$ therefore agree with Bayesian credible intervals based on a uniform prior. It follows that the confidence interval estimator C avoids uniform loss.

Moreover, Buehler (1959) proved that there are no relevant subsets for C .²³ However, the next example shows that C can incur sure loss. Thus the

non-existence of relevant subsets is necessary but not sufficient to avoid sure loss.

7.5.10 Double exponential distributions

Let the sampling model be the real location family based on the double exponential density $f_0(y) = \frac{1}{2}\exp(-|y|)$. Let C be the event that $x \geq \theta$, so $P(C|\theta) = \frac{1}{2}$ for all θ . This defines an optimal (uniformly most accurate), one-sided, 50% confidence interval estimator for θ .

Let B be the event that x is negative. Intuitively, if B occurs then it is less likely that $x \geq \theta$, so we can expect the coverage probabilities conditional on B to be less than $\frac{1}{2}$. Formally, when $\theta > 0$, $P(B(C - \frac{1}{2})|\theta) = -P(B|\theta)/2 = -\exp(-\theta)/4$. When $\theta \leq 0$, $P(B(C - \frac{1}{2})|\theta) = -P(B^c(C - \frac{1}{2})|\theta) = -P(B^c|\theta)/2 = -\exp(\theta)/4$. So $P(-B(C - \frac{1}{2})|\theta) = \exp(-|\theta|)/4$ is positive for all real θ . It follows that the conditional coverage probabilities $P(C|B;\theta)$ are less than the unconditional probability $\frac{1}{2}$ for all θ . B is not a relevant subset, however, as $P(C|B;\theta) \rightarrow \frac{1}{2}$ as $\theta \rightarrow -\infty$.

Nevertheless, we can construct a sure loss by choosing Z in I2 so that $P(-B(C - \frac{1}{2}) - Z|\theta)$ remains positive for all θ . This is achieved by defining $Z(x) = \exp(2x)/2$ when $x < 0$, $Z(x) = 0$ otherwise.²⁴ Then I2 fails, and C incurs sure loss. The biased betting procedure is to bet whenever $x < 0$ that $C(x)$ does not cover θ , at rate $(1 - \exp(2x))/2$.²⁵

7.5.11 Uniform distributions

There can be a sure loss even for a single observation from a location family, and things can be expected to get worse in larger samples. To illustrate that, consider a sample of n independent observations x_1, \dots, x_n from the uniform distribution on the interval $[\theta - 1, \theta + 1]$. Denote the minimum of the observations by $\alpha(x)$ and the maximum by $\beta(x)$. The optimal (uniformly most accurate) one-sided confidence intervals for θ are $C(x) = [\max\{\alpha(x) + 1 - \kappa, \beta(x) - 1\}, \alpha(x) + 1]$, with confidence coefficient $\gamma = 1 - (1 - \kappa/2)^n$, so that $0 < \kappa < 2$ when $0 < \gamma < 1$.²⁶

Consider first the case of a single observation, $n = 1$. Because this is a location family there are no relevant subsets and no uniform loss. But there is a sure loss. To see that, consider the case $\gamma = 0.5$ so that $C(x) = [x, x + 1]$. Let Y be any strictly decreasing, bounded function on the real line, and $Z(x) = \min\{Y(x) - Y(x + 1), Y(x - 1) - Y(x)\}/8$, so $Z(x)$ is positive for all real x . Verify that $P(Y(C - 0.5)|\theta) = \frac{1}{4} \int_{\theta}^{\theta+1} [Y(x - 1) - Y(x)] dx$ and $P(Z|\theta) \leq \frac{1}{8} \int_{\theta}^{\theta+1} [Y(x - 1) - Y(x)] dx$. Hence $P(Y(C - 0.5) - Z|\theta)$ is positive for all real θ . Thus C incurs sure loss when $n = 1$.

Suppose next that $n \geq 2$. Define $B = \{x: \beta(x) - \alpha(x) \geq 2 - \kappa\}$. Conditional

7.5 CONFIDENCE INTERVALS AND RELEVANT SUBSETS

on B , $C(x) = [\beta(x) - 1, \alpha(x) + 1]$ has coverage probability one under each θ . Hence $P(B(C - \gamma)|\theta) = (1 - \gamma) P(B|\theta) > 0$, where $P(B|\theta)$ does not depend on θ , so that I4 fails and there is a uniform sure loss. Here B is a super-relevant subset.

The event B is also a super-relevant subset, when $n \geq 2$, for the standard two-sided confidence intervals.²⁷ These also incur uniform sure loss. In fact, when n is sufficiently large and the range of observations $\beta - \alpha$ is sufficiently small, the two-sided confidence interval can be empty! (For 90% intervals, this can happen whenever n is at least 5.)

7.5.12 Example of uniform loss

In each of the previous examples the estimator C incurs sure loss. We now show, through an artificial example, that C can avoid sure loss but still incur uniform loss. Consider the sampling models with probabilities $P(x|\theta)$ as follows.

$P(x \theta)$	x_1	x_2	x_3	x_4
θ_1	0.1	0.1	0.8	0
θ_2	0.1	0.1	0	0.8
θ_3	0	0	0.9	0.1
θ_4	0	0	0.1	0.9

Define a 90% confidence set estimator C by $C(x_1) = \{\theta_1\}$, $C(x_2) = \{\theta_2\}$, $C(x_3) = \{\theta_1, \theta_3\}$, $C(x_4) = \{\theta_2, \theta_4\}$. This avoids sure loss, because for any gamble Y , $P(Y(C - 0.9)|\theta_3) = 0.09(Y(x_3) - Y(x_4))$ and $P(Y(C - 0.9)|\theta_4) = 0.09(Y(x_4) - Y(x_3))$, and these cannot both be positive.

But there is a uniform loss. To see that, consider the gamble $Y = -\{x_1, x_2\}$. In Theorem 7.5.4, $\tilde{J}(Y) = J(Y) = \{\theta_1, \theta_2\}$. We compute $P(Y(C - 0.9)|\theta_1) = P(Y(C - 0.9)|\theta_2) = 0.08$, so that I3 fails.²⁸

The loss here is uniform on $\{\theta_1, \theta_2\}$. If we reduced the parameter space to $\{\theta_1, \theta_2\}$ it would be a uniform sure loss. But if θ_3 and θ_4 are possible then there can be no sure loss, irrespective of how $P(\cdot|\theta)$ and C are defined for other values of θ . The definitions here are intuitively inconsistent since x_1 supports θ_1 and θ_2 equally strongly, and so does x_2 , yet $C(x_1) = \{\theta_1\}$ and $C(x_2) = \{\theta_2\}$ are 90% confidence sets.²⁹ This again indicates that avoiding sure loss is too weak a requirement.

7.5.13 Other interpretations of confidence intervals

It is clear from these examples that the confidence coefficient γ cannot be coherently interpreted as a precise posterior probability that the confidence

interval covers θ . Even in simple problems involving location or scale parameters, such probabilities incur sure loss.³⁰ It remains to consider whether some alternative interpretation of γ is viable.

One alternative is to interpret γ as a posterior lower probability of coverage, so that $P(C|x) = \gamma$.³¹ On this interpretation, You should be willing to bet on the event that $C(x)$ contains θ at rates arbitrarily close to γ , but not necessarily willing to bet against this event. Your posterior ‘degree of confidence’ is at least γ .

The general conditions in section 7.1 still apply, and Theorem 7.5.4 can be extended to characterize coherence of the assessments $P(C|\mathcal{X}) = \gamma$ with $P(\cdot|\Theta)$. In order to produce a sure (or partial or uniform) loss in 7.5.4, we must be able to choose Y to be non-positive, so that $P(C - \gamma|x) = 0$ implies $P(Y(C - \gamma)|x) = 0$. So the four axioms in 7.5.4 need to be modified by adding the restriction $Y \leq 0$. For example, C incurs sure loss whenever there is a negatively biased relevant subset B , by taking $Y = -B$, but not necessarily when the relevant subsets are positively biased.

Even under this weaker interpretation, optimal confidence intervals often produce a sure loss, e.g. in the double exponential location family (7.5.10).³² Similarly, there is still a uniform sure loss in Examples 7.5.3 and 7.5.11 (if $n \geq 2$), a sure loss in 7.5.2 and 7.5.11 (if $n = 1$), and a uniform loss in 7.5.12. So it does not seem viable to interpret γ as a posterior lower probability.

Neyman’s own interpretation of confidence intervals is based on his concept of ‘inductive behaviour’. He argues that You should ‘assert (or act on the assumption) that’ θ belongs to $C(x)$, on the grounds that, in the long run, such assertions will ‘be frequently correct, and this irrespective of the value that θ may possess’.³³ If You continue to use confidence intervals with confidence coefficient γ , the long-run relative frequency of correct assertions will be γ . But how does that justify a particular inference described by $(C(x), \gamma)$? When $C(x)$ is empty, as in Examples 7.5.2, 7.5.3 and 7.5.11, it seems absurd to ‘assert (or act on the assumption) that’ θ belongs to $C(x)$. More generally, in order to use $C(x)$ for either inference or decision, it seems necessary to measure the posterior uncertainty about whether $C(x)$ covers θ . We have seen that γ is not a suitable measure.

A third interpretation of confidence intervals has been suggested by Kempthorne and Folks (1971) and Cox and Hinkley (1974).³⁴ This relies on an ordering of the possible observations according to their consistency with θ . Define the **consistency level** of θ with x as the probability under $P(\cdot|\theta)$ of observing data that are no more consistent than x with θ . Then $C(x)$ is the set of parameter values whose consistency level with x is at least $1 - \gamma$. When $C(x)$ is empty, that simply means that none of the hypothetical sampling models is consistent with x at the $1 - \gamma$ level.

This interpretation seems more promising than the earlier ones, but it still does not tell us how confident we should be that $C(x)$ covers θ . It is not

clear how consistency levels, which depend on the probabilities assigned to unobserved data, are relevant after observing x . Moreover, there are fundamental problems in ordering the possible observations according to their consistency with θ , and in deciding whether consistency levels should be conditional on subsets of the sample space (e.g. relevant subsets).

A final approach, due to Kiefer (1977),³⁵ is to recognize that the confidence coefficient γ is not an adequate measure of posterior ‘confidence’, and instead define a measure $\gamma(x)$ that depends on x . (It is clear from the earlier examples that an adequate measure of confidence must depend on x .) Provided γ is chosen to satisfy $P(C|\theta) \geq P(\gamma|\theta)$ for all $\theta \in \Theta$, the confidence interval estimator retains the frequentist interpretation that, in repeated trials, the long-run relative frequency with which the sets $C(x)$ cover θ will be at least as large as the average of the reported measures $\gamma(x)$. Coherence of the posterior probabilities $P(C|x) = \gamma(x)$ with $P(\cdot|\Theta)$ is still characterized by the four axioms in Theorem 7.5.4, provided γ is regarded as a function of x .

The basic problem in this approach is to choose the function γ . One way to do so is to choose a partition \mathcal{B} of \mathcal{X} , and define, for each $B \in \mathcal{B}$ and $x \in B$, $\gamma(x) = \inf\{P(C|B; \theta) : \theta \in \Theta\}$. By the conglomerative property, $P(C|\theta) = P(P(C|\mathcal{B}; \theta)|\theta) \geq P(\gamma|\theta)$, so the frequentist interpretation is retained. The partition \mathcal{B} can sometimes be chosen to consist of relevant subsets, e.g. $\mathcal{B} = \{B, B^c\}$ in Example 7.5.3 yields the correct measure of confidence, $\gamma(x) = 1$ for $x \in B$ and $\gamma(x) = 0$ for $x \in B^c$. But there may be several different kinds of relevant subsets, and there does not seem to be any general method of choosing a reasonable \mathcal{B} or γ .³⁶

From a frequentist point of view, there are strong reasons to require, at least, that a confidence interval estimator C has no super-relevant gambles (axiom I4). But that implies that the confidence intervals can be regarded as Bayesian credible intervals, and $\gamma(x)$ as the posterior probability of coverage, resulting from some prior prevision. By Theorem 7.5.4(iv), the posterior probabilities $P(C|x) = \gamma(x)$ are weakly coherent with the sampling model, and it follows (using Theorem 8.2.5) that they are also weakly coherent with a precise prior prevision P on Θ .³⁷ Provided x has positive prior probability, $\gamma(x)$ must satisfy Bayes’ rule, $\gamma(x) = P(L_x C(x))/P(L_x)$, where L_x is the observed likelihood function. More generally, the posterior lower probabilities $P(C|x) = \gamma(x)$ must be weakly coherent with a prior lower prevision P .³⁸ It is therefore natural to define posterior probabilities $\gamma(x)$ by choosing a prior prevision P or P , although this will not usually satisfy the frequentist requirement that $P(C|\theta) \geq P(\gamma|\theta)$ for all θ .³⁹

7.5.14 How much confidence can we have in confidence intervals?

There are two fundamental problems that plague all the interpretations of confidence intervals, and indeed all frequentist theories of statistics. The

first problem is to select an appropriate conditioning set B in which to evaluate coverage probabilities $P(C|B; \theta)$. (The examples show that some kind of conditioning will often be necessary.) In order to rule out relevant subsets, it seems that the conditioning set must be ‘small’, but generally that is incompatible with the frequentist requirement that $P(C|B; \Theta)$ be bounded away from zero. The preceding argument shows, in fact, that all well-behaved confidence intervals (for which there are no super-relevant gambles) can be obtained as Bayesian credible intervals, by conditioning on the observation x ! There appear to be relevant subsets or gambles for most standard confidence intervals. In that case, the frequentist justification of confidence intervals, that they have coverage probability γ under every possible state θ , has little force.

The second problem is to explain how a confidence interval is relevant or useful after observing x . We have seen that γ cannot be interpreted as a posterior probability of coverage, and it is doubtful that γ is an appropriate measure of posterior ‘confidence’ or ‘consistency’ in any other sense. The basic question, ‘how much confidence can we have in $C(x)$ after observing x ?’, remains unanswered. It is therefore unclear how the inference represented by $C(x)$ and γ can be useful for drawing conclusions about θ , guiding further inquiry, or making decisions. In our view, the Neyman–Pearson approach is misconceived.⁴⁰

7.6 Proper prior previsions

The preceding sections were concerned with coherence of sampling models $P(\cdot|\Theta)$ and posterior previsions $P(\cdot|\mathcal{X})$. Whenever these are coherent, they are coherent also with some prior prevision \underline{P} . In statistical problems it will usually be helpful to assess a prior prevision. It then becomes necessary to examine coherence of the three specified previsions P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$. That problem is considered next. In this section \underline{P} is defined only on Θ -measurable gambles and represents beliefs about the parameter value prior to the statistical experiment.¹

We adopt assumptions 7.3.2, so that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are defined on the same domain \mathcal{F} of measurable gambles. The extra assumptions needed here are as follows.

7.6.1 Assumptions

Let \mathcal{K} be the linear subspace of $\mathcal{L}(\Theta \times \mathcal{X})$ consisting of all \mathcal{C} -measurable gambles, where \mathcal{C} is the σ -field of subsets of Θ in 7.3.2. (So all gambles in \mathcal{K} are functions only of θ , and \mathcal{K} is contained in \mathcal{F} .) Assume that \underline{P} is a coherent unconditional lower prevision defined on domain \mathcal{K} .

We will use the following consequences of these assumptions. (As usual,

7.6 PROPER PRIOR PREVISIONS

$G(X)$ denotes the marginal gamble $X - \underline{P}(X)$, and $S(Y)$ is the Θ -support of Y .)

- (a) If $X \in \mathcal{K}$ then $G(X) \in \mathcal{K}$ and $\underline{P}(G(X)) = 0$.
- (b) If $Y \in \mathcal{F}$ then $\underline{P}(Y|\Theta) \in \mathcal{K}$ and $S(Y) \in \mathcal{K}$ (using 7.3.2).

In terms of the notation in section 7.1, we have three previsions $\underline{P}(\cdot|\mathcal{B}_1) = P$, $\underline{P}(\cdot|\mathcal{B}_2) = P(\cdot|\Theta)$, $\underline{P}(\cdot|\mathcal{B}_3) = P(\cdot|\mathcal{X})$. Theorem 7.1.5 shows that the three previsions avoid partial loss if and only if (a) $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ avoid partial loss, and (b) P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ avoid uniform sure loss. This remains true if ‘partial’ is replaced by ‘uniform’ or ‘sure’. We have characterized the conditions for (a) in section 7.3. The extra rationality conditions imposed when we introduce \underline{P} are the same whichever of the general requirements 7.1.1–7.1.3 we adopt. The extra condition (b) can be written as the following axiom S6.

7.6.2 Uniform sure loss theorem

Suppose that P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ satisfy assumptions 7.3.2 and 7.6.1. They avoid uniform sure loss if and only if they satisfy the axiom

$$(S6) \quad \underline{P}(P(Y|\Theta)) \leq \sup \bar{P}(Y|\mathcal{X}) \quad \text{for all } Y \in \mathcal{F}.$$

Proof. By 7.1.1, avoiding uniform sure loss is equivalent to

$$\sup U \geq 0 \text{ whenever } U = G(X) + G(V|\Theta) + G(W|\mathcal{X}), X \in \mathcal{K}, V \in \mathcal{F}, W \in \mathcal{F}.$$

This implies S6 by taking $X = \underline{P}(Y|\Theta)$, $V = Y$, $W = -Y$ to give $U = \bar{P}(Y|\mathcal{X}) - \underline{P}(P(Y|\Theta))$.

Conversely, suppose S6 holds and X, V, W are given. Let $Y = G(X) + G(V|\Theta)$, so $Y \in \mathcal{F}$. Then $\underline{P}(Y|\Theta) = G(X)$ by separate coherence of $\underline{P}(\cdot|\Theta)$, and $\underline{P}(\underline{P}(Y|\Theta)) = 0$ by coherence of \underline{P} . Also $Y = U - G(W|\mathcal{X}) \leq \sup U - G(W|\mathcal{X})$, so $\bar{P}(Y|\mathcal{X}) \leq \sup U$ by separate coherence of $P(\cdot|\mathcal{X})$. Applying S6, $\sup U \geq \sup \bar{P}(Y|\mathcal{X}) \geq \underline{P}(P(Y|\Theta)) = 0$. Thus the three previsions do avoid uniform sure loss. ◆

7.6.3 Corollary

Under assumptions 7.3.2 and 7.6.1, P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ avoid partial loss if and only if they satisfy axioms S1 and S6. (For avoiding sure loss or avoiding uniform loss, replace S1 by S2 or S3 respectively.)

Proof. Apply Theorems 7.1.5 and 7.3.3. ◆

Compare the new axiom S6 with axiom C7 in section 6.5.2. Defining E on domain \mathcal{F} as the natural extension of \underline{P} and $P(\cdot|\Theta)$, by $E(Y) = \underline{P}(P(Y|\Theta))$, S6 is simply the requirement (C7) that E and $P(\cdot|\mathcal{X})$

avoid sure loss.² By Theorem 6.5.3, this is equivalent to axiom C10, and S6 may therefore be rewritten as the equivalent axioms

$$(S6a) \bar{E}(G(Y|\mathcal{X})) \geq 0 \text{ for all } Y \in \mathcal{F}$$

$$(S6b) \bar{E}(Y) \geq \underline{E}(\underline{P}(Y|\mathcal{X})) \text{ for all } Y \in \mathcal{F}$$

$$(S6c) \underline{E}(Y) \leq \bar{E}(\bar{P}(Y|\mathcal{X})) \text{ for all } Y \in \mathcal{F}.$$

When all the previsions P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are linear, as in standard Bayesian inference, S6 may be written as

$$(S6^*) E(Y) = E(P(Y|\mathcal{X})) \text{ for all } Y \in \mathcal{F}.$$

Consider next the relationship between S6 and the axioms S1–S4 concerning coherence of $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. Since $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ avoid uniform sure loss if and only if they satisfy S4, the last theorem shows that S6 implies S4. If every singleton $\{\theta\}$ has positive prior lower probability, or if every possible observation has positive prior lower probability, then S6 implies S1. Thus, when either Θ or \mathcal{X} is discrete, avoiding partial loss will often reduce to axiom S6.

In general, however, S6 does not imply S1, nor even the weaker axioms S2 and S3. To see that S6 does not imply S3, extend Example 7.3.5 by defining the prior prevision P to be degenerate at $\theta = 3$. It is easily seen that S6 is satisfied, but S3 fails.⁴

The next example shows that S6 is not sufficient for S2.

7.6.4 Normal distributions

As in section 7.4.4, let the sampling distributions be Normal $(\theta, 1)$ and let the posterior distributions be Normal $(x, 1)$. Let P be a translation-invariant linear prevision on $\Theta = \mathbb{R}$. Heath and Sudderth (1978) prove that such P exist and that P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ satisfy S6*, so they avoid uniform sure loss. But we saw in section 7.4.4 that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ do not avoid sure loss. So S6 does not imply S2.⁵

It is easy to see that S6 does imply S2 when both P and $P(\cdot|\theta)$ are countably additive linear previsions, so that their natural extension E is also. If also all the posteriors are absolutely continuous with respect to the prior then S6 implies S1.⁶

Now consider the coherence condition 7.1.4. When all the previsions P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are linear, they are coherent if and only if they satisfy S1 and S6*. More generally we can characterize coherence through the next axiom S7. We use the notation $\bar{P}_{x_0}(Y|\mathcal{X})$ introduced in section 6.5.1, and similarly define $\underline{P}_{\theta_0}(Y|\Theta) = \{\theta_0\}\bar{P}(Y|\theta_0) + \{\theta_0\}^c\underline{P}(Y|\Theta)$. These gambles belong to \mathcal{X} , provided Y is in \mathcal{F} .

7.6.5 Coherence theorem⁷

Suppose the previsions \underline{P} , $\underline{P}(\cdot|\Theta)$, $\underline{P}(\cdot|\mathcal{X})$ satisfy assumptions 7.3.2 and 7.6.1. The previsions are weakly coherent if and only if they satisfy the three-part axiom

- (S7) (i) $\bar{P}(\underline{P}(Y|\Theta)) \leq \sup \bar{P}(Y|\mathcal{X}) \quad \text{for all } Y \in \mathcal{F}$
- (ii) $\underline{P}(\underline{P}_{\theta_0}(Y|\Theta)) \leq \sup \bar{P}(Y|\mathcal{X}) \quad \text{for all } Y \in \mathcal{F} \text{ and } \theta_0 \in \Theta$
- (iii) $\underline{P}(\underline{P}(Y|\Theta)) \leq \sup \bar{P}_{x_0}(Y|\mathcal{X}) \quad \text{for all } Y \in \mathcal{F} \text{ and } x_0 \in \mathcal{X}.$ ⁸

The previsions are coherent if and only if they satisfy axioms S5 and S7.

Each of the three parts of S7 implies S6. In fact, the three parts say that S6 continues to hold when we modify a single prevision, by (i) replacing the prior \underline{P} by \bar{P} , (ii) replacing the sampling model $\underline{P}(\cdot|\theta_0)$ by $\bar{P}(\cdot|\theta_0)$, or (iii) replacing the posterior $\bar{P}(\cdot|x_0)$ by $\underline{P}(\cdot|x_0)$.

To give examples of coherent non-linear previsions, we show next that it is coherent to adopt vacuous posterior beliefs whenever either Your prior beliefs are vacuous, or the observation x has prior probability zero. In both cases there is no information contained in Your prior previsions for gambles contingent on x from which to construct a non-vacuous posterior.

7.6.6 Vacuous prior

Let \underline{P} be the vacuous prior prevision, $\bar{P}(\cdot|\mathcal{X})$ the vacuous posterior previsions, and $\underline{P}(\cdot|\Theta)$ any separately coherent sampling model. Axiom S5 is satisfied, by Example 7.3.7. To verify that S7 also holds, note that $\sup \bar{P}(Y|\mathcal{X}) = \sup Y$, from which (i) and (ii) follow. For S7 (iii), find θ_0 such that $Y(\theta_0, x_0) - \delta \leq \inf Y(\cdot, x_0) = \underline{P}(Y|x_0)$. Then $\bar{P}_{x_0}(Y|x) + \delta \geq Y(\theta_0, x)$ for all x , hence

$$\sup \bar{P}_{x_0}(Y|\mathcal{X}) + \delta \geq \underline{P}(Y|\theta_0) \geq \inf \underline{P}(Y|\Theta) = \underline{P}(\underline{P}(Y|\Theta)).$$

Thus it is coherent, for any sampling model, to hold prior and posterior beliefs about Θ that are both vacuous. Indeed, stronger results are established in sections 7.4.1, 8.2 and 8.4.7: non-vacuous posterior previsions typically imply non-vacuous prior previsions, by natural extension. When Your prior beliefs are vacuous, coherence requires Your posterior beliefs to be vacuous. On the other hand, the next example shows that vacuous posterior previsions can be coherent with precise prior previsions.

7.6.7 Observations with zero prior probability

Let $\underline{P}(\cdot|\mathcal{X})$ be the vacuous posterior previsions, and let \underline{P} and $\underline{P}(\cdot|\Theta)$ be any separately coherent previsions such that the prior probability $\underline{P}(\bar{P}(\{x\}|\Theta))$

is zero for every x in \mathcal{X} . Then \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent. (To see that, use Example 7.3.7 to establish S5. Parts (i) and (ii) of S7 hold as in the previous example, and (iii) holds using the zero prior probabilities by result 4 of section 7.6.8.) Thus, when all observations have zero prior probability it is coherent to adopt vacuous posterior beliefs.

The coherence axioms S5 and S7 are somewhat more complicated than the avoiding loss axioms S1 and S6, but they simplify in the following important cases.

7.6.8 Simplifications of coherence

Suppose that \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ satisfy assumptions 7.3.2 and 7.6.1. Let \underline{E} denote the natural extension of \underline{P} and $\underline{P}(\cdot|\Theta)$, defined by $\underline{E}(Y) = \underline{P}(P(Y|\Theta))$. Then the coherence axioms S5 and S7 simplify as follows.

1. If $\underline{P}(\{\theta\}) > 0$ for every $\theta \in \Theta$ then S7 implies S5.
2. If $\underline{E}(\{x\}) > 0$ for every $x \in \mathcal{X}$ then S7 implies S5. (So, when either Θ or \mathcal{X} is a countable set, coherence will often reduce to axiom S7.)
3. If $\underline{P}(\{\theta\}) = 0$ for every $\theta \in \Theta$ then S7(i) implies S7(ii).
4. If $\underline{P}(\underline{P}(\{x\}|\Theta)) = 0$ for every $x \in \mathcal{X}$ then S7(i) implies S7(iii).⁹
5. If $\bar{\underline{P}}(\{\theta\}|\mathcal{X}) = 0$ for every $\theta \in \Theta$ then S1 implies S5(i).
6. If $\bar{\underline{P}}(\{x\}|\Theta) = 0$ for every $x \in \mathcal{X}$ then S1 implies S5(ii).
7. When both spaces Θ and \mathcal{X} are uncountable we will typically have $\underline{P}(\{\theta\}) = 0$, $\bar{\underline{P}}(\{\theta\}|x) = 0$ and $\bar{\underline{P}}(\{x\}|\theta) = 0$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. In that case, by the preceding results, S5 reduces to S1 and S7 reduces to S7(i). Coherence is then equivalent to S1 and S7(i).¹⁰
8. Suppose that the sampling models $P(\cdot|\theta)$ are linear previsions, as in standard statistical problems. Then S5(i) reduces to S1 and S7(ii) reduces to S6. Also (i) and (iii) of S7 are equivalent to axioms C8 and C9 (or C11 and C12) applied to \underline{E} and $\underline{P}(\cdot|\mathcal{X})$. Thus coherence reduces to coherence of the pair $P(\cdot|\Theta)$, $\underline{P}(\cdot|\mathcal{X})$, plus coherence of the pair $\underline{E}, \underline{P}(\cdot|\mathcal{X})$.¹¹ Whenever $\underline{E}(\{x\}) > 0$, the posterior $\underline{P}(\cdot|x)$ is uniquely determined by \underline{P} and $P(\cdot|\Theta)$, through $\underline{E}(G(Y|x)) = 0$ for all $Y \in \mathcal{F}$. (This is the generalized Bayes rule, C12.)
9. If the posterior previsions are all linear then S5(ii) reduces to S1 and S7(iii) reduces to S6.
10. If both $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are linear, as for inferences from improper priors and the other methods discussed in section 7.4, coherence reduces to axioms S1 and S7(i).
11. If the prior P is linear then S7(i) reduces to S6.

12. If all the previsions P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are linear, as in Bayesian inference, then coherence is equivalent to axioms S1 and S6*. Standard Bayesian models are discussed in the next section.

7.7 Standard Bayesian inference

By standard Bayesian inferences we mean inferences in which the prior prevision is linear and countably additive, and posterior densities are defined through Bayes' rule. The main result of this section is that, under the usual regularity conditions, standard Bayesian inferences are coherent.

7.7.1 Definition

Say that P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are a **standard Bayesian model** when they satisfy the following conditions. (As in assumptions 7.3.2, \mathcal{C} and \mathcal{S} are σ -fields of subsets of Θ and \mathcal{X} respectively.)

- (a) P is a countably additive, linear prevision defined on the domain \mathcal{F} of all \mathcal{C} -measurable gambles. (So its restriction to \mathcal{C} is a probability measure.)
- (b) The domain \mathcal{F} of both $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ is the linear space of all $\mathcal{C} \times \mathcal{S}$ -measurable gambles. For each $\theta \in \Theta$, $P(\cdot|\theta)$ is a countably additive, linear prevision on \mathcal{F} and has density function $f(\cdot|\theta)$ with respect to a σ -finite measure v on \mathcal{S} , so that $P(Y|\theta) = \int Y(\theta, x)f(x|\theta)v(dx)$ for all $Y \in \mathcal{F}$ and $\theta \in \Theta$.¹ We assume that $f(x|\theta)$ is $\mathcal{C} \times \mathcal{S}$ -measurable as a function of (θ, x) , so that $P(Y|\Theta)$ is a \mathcal{C} -measurable gamble whenever $Y \in \mathcal{F}$.
- (c) For each $X \in \mathcal{X}$, define $q(x) = \int f(x|\theta)P(d\theta)$. (This may be interpreted as the predictive density function of the observation with respect to v .) Then q is \mathcal{S} -measurable and non-negative.² For all x such that $0 < q(x) < \infty$, define the **standard Bayesian posterior density** $g(\cdot|x)$ by Bayes' rule $g(\theta|x) = f(x|\theta)/q(x)$. Then $g(\cdot|x)$ is a probability density function with respect to P . When $q(x) = 0$ or $q(x) = \infty$, $g(\cdot|x)$ can be taken to be any probability density function with respect to P , provided $g(\theta|x)$ is $\mathcal{C} \times \mathcal{S}$ -measurable as a function of (θ, x) to ensure that $P(Y|\mathcal{X})$ is \mathcal{S} -measurable whenever $Y \in \mathcal{F}$.³ Define the posterior previsions $P(\cdot|\mathcal{X})$ by $P(Y|x) = \int Y(\theta, x)g(\theta|x)P(d\theta)$ for all $Y \in \mathcal{F}$ and $x \in \mathcal{X}$.

Each posterior prevision $P(\cdot|x)$ is a countably additive, linear prevision on \mathcal{F} and has density $g(\cdot|x)$ with respect to P . Thus $g(\cdot|x)$ and $P(\cdot|x)$ are

uniquely determined by the prior P and sampling densities $f(x|\theta)$ unless $q(x)$ is zero or infinite (which event has prior probability zero).

7.7.2 Bayesian coherence theorem

If $P, P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are a standard Bayesian model then they are coherent.

Proof. The definition of a standard Bayesian model ensures that assumptions 7.3.2 and 7.6.1 are satisfied. Let $E(Y) = P(P(Y|\Theta))$ denote the natural extension of P and $P(\cdot|\Theta)$ to \mathcal{F} . By Corollary 7.6.3, coherence reduces to the two axioms

(S1) there is no non-zero Y in \mathcal{F} such that $P(Y|\theta) > 0$ for all $\theta \in S(Y)$ and $P(Y|x) < 0$ for all $x \in T(Y)$,

(S6*) $E(Y) = E(P(Y|\mathcal{X}))$ for all $Y \in \mathcal{F}$.

Consider S6* first. For any $Z \in \mathcal{F}$, $E(Z) = \int P(Z|\theta)P(d\theta) = \iint Z(\theta, x)f(x|\theta)v(dx)P(d\theta)$. (The integrand here is $\mathcal{C} \times \mathcal{S}$ -measurable because Z and $f(\cdot|\cdot)$ are.) We can use Fubini's theorem to change the order of integration because $E(|Z|)$ is finite when Z is bounded.

For $A \subset \mathcal{X}$ we have $E(A) = \int_A \int_{\Theta} f(x|\theta)P(d\theta)v(dx) = \int_A q(x)v(dx)$. When A is the event that $q(x) = 0$ we get $E(A) = 0$. When A is the event that $q(x) = \infty$ we have $v(A) = 0$ so $E(A) = 0$. Taking B to be the event that $0 < q(x) < \infty$, it follows that $E(B) = 1$ and $E(Z) = E(BZ)$ for all $Z \in \mathcal{F}$.

Apply these results to $Z = P(Y|\mathcal{X})$, and use the definition of $P(Y|x)$ and Fubini's theorem to show

$$\begin{aligned} E(P(Y|\mathcal{X})) &= E(BP(Y|\mathcal{X})) = \int_B \int_{\Theta} f(x|\theta)P(d\theta)P(Y|x)v(dx) \\ &= \int_B q(x)P(Y|x)v(dx) = \int_B q(x) \int_{\Theta} Y(\theta, x)g(\theta|x)P(d\theta)v(dx) \\ &= \int_B \int_{\Theta} Y(\theta, x)f(x|\theta)P(d\theta)v(dx) \\ &= \int_{\Theta} \int_B Y(\theta, x)f(x|\theta)v(dx)P(d\theta) = E(BY) = E(Y). \end{aligned}$$

This establishes S6*.

Suppose next that S1 fails for some $Y \in \mathcal{F}$. Then $P(Y|\mathcal{X}) \leq 0$, hence $E(Y) = E(P(Y|\mathcal{X})) \leq 0$ by S6*. If $P(S(Y)) > 0$ then $E(Y) = P(P(Y|\Theta)) = P(S(Y)P(Y|\Theta)) > 0$ (a contradiction), using countable additivity of P . If $P(S(Y)) = 0$ then $P(Y|x) = \int Y(\theta, x)g(\theta|x)P(d\theta) = \int_{S(Y)} Y(\theta, x)g(\theta|x)P(d\theta) = 0$ for all $x \in \mathcal{X}$, which is inconsistent with failure of S1. Thus S1 must be satisfied.⁴ ♦

7.7 STANDARD BAYESIAN INFERENCE

Examples of standard Bayesian models are well known. We mention only the following.

7.7.3 Beta-binomial model

Here $\Theta = [0, 1]$, $\mathcal{X} = \{0, 1, \dots, n\}$, \mathcal{C} is the Borel σ -field of subsets of Θ , \mathcal{S} contains all subsets of \mathcal{X} , so \mathcal{F} is the class of all gambles Y such that $Y(\cdot, x)$ is Borel-measurable for each $x \in \mathcal{X}$. Take v to be counting measure on \mathcal{S} , so the density $f(x|\theta)$ is just the probability $P(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, which defines the binomial (n, θ) sampling distribution. The prior P is a beta (s, t) distribution, which has density $h(\theta) \propto \theta^{s-1}(1-\theta)^{t-1}$ with respect to Lebesgue measure, where $0 < t < 1$ and $s > 0$.

Here $q(x) = \int_0^1 f(x|\theta)h(\theta)d\theta$ satisfies $0 < q(x) < \infty$ for all $x \in \mathcal{X}$. For all x , the posterior prevision has density $f(x|\theta)/q(x)$ with respect to P , and density $h(\theta)f(x|\theta)/q(x)$ with respect to Lebesgue measure. The latter density is proportional to $\theta^{s-1+x}(1-\theta)^{t-1+n-x}$ as a function of θ , and this is recognized as the density of a beta (s', t') distribution where $s' = s + n$, $t' = (st + x)/(s + n)$. These posterior beta distributions are coherent with the beta (s, t) prior and binomial sampling model.⁵

7.7.4 Uniqueness of posterior

Suppose that the prior P and sampling model $P(\cdot|\Theta)$ satisfy (a) and (b) of the standard Bayesian model (7.7.1), and that $0 < q(x) < \infty$ for all x . (Otherwise the posterior can be indeterminate.) Then the standard Bayesian posteriors, defined by (c), are coherent with P and $P(\cdot|\Theta)$. Can there be other coherent posteriors? The standard Bayesian posteriors are the unique coherent posteriors provided every x has positive probability $E(\{x\})$, since then the posterior is uniquely determined through Bayes' rule. More generally, a linear posterior $P(\cdot|\mathcal{X})$ that is coherent with P and $P(\cdot|\Theta)$ must be almost unique, in the sense that axiom S6* determines the values $P(Y|\mathcal{X})$ up to a subset of \mathcal{X} with prior probability zero. However, coherence is preserved if we modify the standard Bayesian densities $g(\cdot|x)$ for all x in any set of v -measure zero. When the sample space is continuous, the observation x will usually have prior probability zero, and the posterior $P(\cdot|x)$ is entirely unconstrained by P and $P(\cdot|\Theta)$.

One possible justification for the standard Bayesian posterior is to regard the points in a continuous space \mathcal{X} as idealizations of discrete events $B(x, \delta)$. As in Theorem 6.10.4, the posteriors $P(\cdot|x)$ may be determined as limits of coherent posteriors $P(\cdot|B(x, \delta))$. However, this requires information about the idealizations involved in constructing the continuous model, and cannot be justified merely on grounds of mathematical regularity or naturalness.

Borel's paradox (6.10.1) shows that several different types of idealization, leading to different posteriors, may be 'natural' in a specific problem. Without further information the prior and sampling model do not determine a unique coherent posterior and this need not be defined as the standard Bayesian posterior.

There is even greater indeterminacy in the posteriors $\underline{P}(\cdot|\mathcal{X})$ for continuous \mathcal{X} if non-linear posteriors are admitted. If every observation x has prior probability zero, which holds when $v(\{x\})=0$ for all x , then the vacuous posteriors are coherent with the prior and sampling model (by Example 7.6.7). A very wide range of non-vacuous but non-linear posteriors will also be coherent.⁶

We conclude that for continuous sample spaces there is nothing compelling about the standard Bayesian posterior or the continuous version of Bayes' rule. For discrete sample spaces, when the observations have positive prior probability, Bayes' rule is a consequence of coherence, and the standard Bayesian posterior is uniquely coherent.

7.7.5 Finitely additive priors

Countable additivity of the prior prevision P was used in Theorem 7.7.2 in three different ways:

1. to define the posterior previsions in terms of densities with respect to P ;
2. to justify changing the order of integration through Fubini's theorem;
3. in establishing S1, to conclude that $E(Y) > 0$ whenever $P(S(Y)) > 0$.

For linear priors that are not countably additive, 1 can be avoided by defining the posterior previsions directly, rather than via Bayes' rule, and 2 can sometimes be justified, thereby establishing S6*.⁷ But it appears that inferences from linear priors that are not countably additive will often violate S1. Such inferences are weakly coherent but not coherent. For example, the translation-invariant prior, Normal $(\theta, 1)$ sampling models and Normal $(x, 1)$ posteriors in Example 7.6.4 satisfy S6*, but we know from section 7.4.4 that they violate S1.⁸

7.7.6 Is it wise to be Bayesian?

Theorem 7.7.2 shows that standard Bayesian inferences are coherent. That is an advantage over non-standard Bayesian inferences from proper priors that are not countably additive. It is a big advantage over inferences from improper priors, and the other incoherent methods mentioned in section 7.4. However, it does not establish that Bayesian inferences are reasonable

7.8 INFERENCES FROM IMPRECISE PRIORS

in practice. We have argued throughout that coherence is a minimal requirement of rationality. Beliefs should be coherent, but they should also conform to evidence. Bayesian inferences rarely conform to evidence. They require precise prior previsions, whereas prior information is rarely adequate to justify precision.⁹

So we doubt that Bayesian inferences (standard or non-standard) are often sensible in practical problems. They may be sensible in the atypical problems where (a) there is a lot of prior information about Θ , justifying a relatively precise prior, or (b) the statistical data provide a lot of information, so that the posterior from a realistic, imprecise prior is quite precise and can be approximated by a Bayesian posterior. More typically there is only a small or moderate amount of information about θ in both the prior evidence and data, and both prior and posterior previsions should be imprecise to reflect this.

We therefore need to develop statistical models that are coherent but imprecise. Several ways of doing so are described in the next section.

7.8 Inferences from imprecise priors

Standard Bayesian inferences, inferences from improper priors, fiducial inferences, and the other objective methods mentioned in section 7.4, all produce precise posterior probabilities. That seems unrealistic when there is little information about Θ . We now describe three ways of constructing imprecise posterior probabilities. The first way is to form the lower envelopes of a class of standard Bayesian priors and the corresponding class of standard Bayesian posteriors. The second is to construct a non-linear prior prevision, and define a non-linear posterior prevision by the generalized Bayes rule, without reference to a class of precise priors. A third method is to define an imprecise posterior without reference to any prior previsions, and check whether it is coherent with the sampling model. We will give examples to illustrate these three methods.

First, Theorem 7.1.6 shows that coherent, imprecise statistical models can be constructed as lower envelopes of standard Bayesian models.

7.8.1 Theorem

Suppose that P_γ , $P_\gamma(\cdot|\Theta)$ and $P_\gamma(\cdot|\mathcal{X})$ are a standard Bayesian model on domains \mathcal{X} and \mathcal{F} , for every γ in a non-empty index set Γ . Then their lower envelopes \underline{P}_γ , $\underline{P}_\gamma(\cdot|\Theta)$ and $\underline{P}_\gamma(\cdot|\mathcal{X})$ are coherent.

Proof. Each standard Bayesian model is coherent by Theorem 7.7.2. Apply Theorem 7.1.6. ◆

Thus standard Bayesian models provide a rich supply of coherent, imprecise inferences. In particular, consider the case where the prior previsions P_γ are all absolutely continuous with respect to a measure ξ , so they have densities h_γ . Provided $q_\gamma(x) = \int f_\gamma(x|\theta)h_\gamma(\theta)\xi(d\theta)$ is finite and non-zero for all γ in Γ ,¹ the posterior densities of $P_\gamma(\cdot|x)$ with respect to ξ are $g_\gamma(\theta|x) = h_\gamma(\theta)f_\gamma(x|\theta)/q_\gamma(x)$. The posterior prevision $\underline{P}(\cdot|x)$ is determined by the functions h_γ and $f_\gamma(x|\cdot)$, by $\underline{P}(Y|x) = \inf\{P_\gamma(Y|x): \gamma \in \Gamma\} = \inf\{\int Y(\theta, x)h_\gamma(\theta)f_\gamma(x|\theta)\xi(d\theta)/q_\gamma(x): \gamma \in \Gamma\}$.

An important case is that in which all the sampling densities f_γ are the same, so the sampling model $P(\cdot|\Theta)$ is linear. An example is the beta-binomial model used in section 5.3.

7.8.2 Beta-binomial model

As in Example 7.7.3, the sampling distributions are binomial (n, θ) and the prior distributions P_γ are beta (s_γ, t_γ) , which are absolutely continuous with respect to Lebesgue measure. The posterior distributions $P_\gamma(\cdot|x)$ are then beta (s'_γ, t'_γ) where $s'_\gamma = s_\gamma + n$, $t'_\gamma = (s_\gamma t_\gamma + x)/(s_\gamma + n)$. These are standard Bayesian inferences. Any class of beta prior distributions indexed by Γ determines a class of beta posterior distributions, and the lower envelopes of the two classes are coherent with the binomial sampling distributions by Theorem 7.8.1. The inferences from several classes Γ were described in sections 5.3 and 5.4.

7.8.3 Normal model

Let $\Theta = \mathcal{X} = \mathbb{R}$, and take \mathcal{K} and \mathcal{F} to consist of Borel-measurable gambles. Let the sampling distributions be Normal $(\theta, 1)$. Consider a class of Normal (μ, η^2) prior distributions, again absolutely continuous with respect to Lebesgue measure. The Bayesian posterior distributions $P_\gamma(\cdot|x)$ are Normal $(\alpha_\gamma(x), \beta_\gamma^2)$ where $\beta_\gamma^{-2} = \eta_\gamma^{-2} + 1$ and $\alpha_\gamma(x) = \mu_\gamma + (x - \mu_\gamma)\beta_\gamma^2$. This defines a standard Bayesian model. Again, any set of prior parameter values (μ, η^2) leads to a set of posterior values $(\alpha(x), \beta^2)$, and the lower envelopes of the corresponding Normal distributions are coherent.

For example, consider the prior class of all Normal $(0, \eta^2)$ distributions with $\eta > c$, where c is a large positive number. This class might represent prior beliefs about θ when You judge that $|\theta|$ may be extremely large, and You regard positive and negative values as equally likely. The posteriors are Normal $(\beta^2 x, \beta^2)$ with $(1 + c^{-2})^{-1} < \beta^2 < 1$. The posterior upper and lower previsions for θ are then x and $x/(1 + c^{-2})$. These are close together, indicating that the posterior prevision $\underline{P}(\cdot|x)$ is relatively precise, just when $|x|$ is much smaller than c^2 .²

7.8 INFERENCES FROM IMPRECISE PRIORS

The improper uniform prior can be regarded as the limit of Normal $(0, \eta^2)$ priors as $\eta^2 \rightarrow \infty$, and the corresponding Normal $(x, 1)$ posterior can be regarded as the limit of Normal $(\beta^2 x, \beta^2)$ posteriors as $\beta^2 \rightarrow 1$. For observations x such that $|x|$ is much smaller than c^2 , the posterior $\underline{P}(\cdot|x)$ is quite precise and almost agrees with the Normal $(x, 1)$ posterior from the improper prior. In this case the inference from the imprecise prior approximates that from the imprecise prior. The approximation gets worse, however, as $|x|$ increases. When $|x|$ is of the same order as c^2 , the class of Normal posteriors still has the Normal $(x, 1)$ posterior as a limit point, but it also contains Normal distributions with quite different means and the posterior $\underline{P}(\cdot|x)$ is quite imprecise. The imprecise model is coherent and its posteriors may be reasonable for all observations x , unlike the inferences from the improper prior.

7.8.4 Imprecise sampling models

Another important case of Theorem 7.8.1 is that in which an imprecise sampling model is specified as a class of linear sampling models $\{P(\cdot|\theta, \psi): \psi \in \Psi\}$, as in section 7.2.7. Suppose that You also assess a class of linear priors $\{P_\gamma: \gamma \in \Gamma\}$ concerning θ . Each pair (γ, ψ) determines a Bayesian posterior $P_{\gamma, \psi}(\cdot|\mathcal{X})$. The posterior prevision $\underline{P}(\cdot|\mathcal{X})$ can then be defined as the lower envelope of Bayesian posteriors over all pairs (γ, ψ) in $\Gamma \times \Psi$. However, other coherent posteriors can be constructed from subsets of $\Gamma \times \Psi$, and may be justified in specific problems. These ideas are illustrated in the following example.³

7.8.5 Imprecise Normal model

Suppose that an unknown physical quantity θ is measured on an instrument which gives reading x , and that the measurement error $x - \theta$ has Normal distribution with mean close to zero and variance close to one. Because the bias v and imprecision σ^2 of the measuring instrument cannot be exactly determined, You might adopt the class of Normal $(\theta + v, \sigma^2)$ distributions, with $|v| \leq \epsilon$ and $\kappa^{-1} \leq \sigma^2 \leq \kappa$, as a robust sampling model for the observation x .⁴ Suppose that Your prior beliefs about θ are represented by a class of Normal (μ, η^2) distributions. Specific values of $(\mu, \eta^2, v, \sigma^2)$ determine posterior Normal $(\alpha(x), \beta^2)$ distributions, where $\beta^{-2} = \eta^{-2} + \sigma^{-2}$, $\alpha(x) = \beta^2(\eta^{-2}\mu + \sigma^{-2}(x - v))$. The posterior $\underline{P}(\cdot|x)$ is the lower envelope of Normal $(\alpha(x), \beta^2)$ distributions over the range of values specified for $(\mu, \eta^2, v, \sigma^2)$.

We consider a simple case of this model for illustration. Assume $\kappa = 1$,

so the measurement variance is exactly one, $\varepsilon = 1$, so the measurement bias is at most one unit, and $\eta^2 = 1$ is constant. Suppose Your prior beliefs about θ are represented by the range $9 \leq \mu \leq 11$. These might be based on a previous measurement $y = 10$, on another instrument whose bias is unknown. If the two instruments are unrelated it is natural to consider all pairs (μ, v) with $9 \leq \mu \leq 11$ and $|v| \leq 1$. This produces the class of Normal $(\alpha + \frac{1}{2}x, \frac{1}{2})$ posteriors over $4 \leq \alpha \leq 6$. The lower envelope of this class, $P_1(\cdot|\mathcal{X})$, is non-linear.

In this model there is no ‘linkage’ between the sampling model and prior; all pairs (μ, v) are used in defining the posterior. The absence of linkage might be justified by the unrelatedness of the two measuring instruments. Suppose instead that prior beliefs about θ are based on a measurement $y = 10$ on an instrument known to have opposite bias $(-v)$ to this one. You might then adopt the same prior class as before, of Normal $(10 + v, 1)$ distributions over $|v| \leq 1$, but these are now linked to the sampling models through the parameter v . Different values of v generate the same Normal $(5 + \frac{1}{2}x, \frac{1}{2})$ posterior distribution. The sampling model and prior are represented by the same classes of Normal distributions as before, but now they generate linear posteriors $P_2(\cdot|\mathcal{X})$.

This example shows that an imprecise sampling model and prior prevision do not determine a unique posterior prevision. Several posteriors $\underline{P}(\cdot|\mathcal{X})$ can be coherent with the same sampling model and prior. One way to determine the posterior is to specify a joint prior prevision for (θ, x) . That leads to the model discussed in Example 7.9.3.

7.8.6 Generalized Bayes rule

Suppose again that the sampling model is linear. One way of constructing imprecise posteriors is to elicit a class of standard Bayesian priors and form the class of Bayesian posteriors. That is the approach followed by Bayesian sensitivity analysts.. But the discussion in Chapter 4 and section 2.10 and 5.9 suggests that it may be simpler to assess a lower prevision \underline{P} directly, rather than as the lower envelope of a class of precise priors.

We may then be able to compute the posterior lower previsions $\underline{P}(\cdot|x)$ directly from \underline{P} , through the **generalized Bayes rule** $\underline{P}((Y - \underline{P}(Y|x))L_x) = 0$, where L_x is the observed likelihood function. When the sample space \mathcal{X} is discrete, $L_x(\theta) = P(\{x\}|\theta)$. When \mathcal{X} is continuous and $P(\cdot|\theta)$ have density functions $f(\cdot|\theta)$ with respect to a common measure, it is usual to define $L_x(\theta) = f(x|\theta)$. Provided $P(L_x) > 0$, the posterior $\underline{P}(\cdot|x)$ is uniquely determined through the GBR.⁵

7.8 INFERENCES FROM IMPRECISE PRIORS

7.8.7 Constant odds-ratio (COR) model

To illustrate this approach, let \underline{P} be the constant odds-ratio model (section 2.9.4) determined by a linear prevision P_0 and constant τ . Suppose that $P_0(L_x) > 0$. Define $\rho = \underline{P}(Y|x)$ through the GBR $\underline{P}((Y - \rho)L_x) = 0$. From 2.9.4, this means that $(1 - \tau)P_0((Y - \rho)^+L_x) = P_0((Y - \rho)^-L_x)$. Dividing by $P_0(L_x)$, and defining $P_0(Z|x) = P_0(ZL_x)/P_0(L_x)$ to be the Bayesian posterior for P_0 , we see that $(1 - \tau)P_0((Y - \rho)^+|x) = P_0((Y - \rho)^-|x)$. This shows that the posterior prevision $\underline{P}(\cdot|x)$ is the COR model determined by $P_0(\cdot|x)$ and τ . Thus the imprecise posterior has the same form as the prior. You need only to update the linear prior P_0 by Bayes’ rule.⁶

7.8.8 Coherence conditions

It is not automatic that the posteriors defined by the GBR will be coherent with the prior and sampling model. By result 8 of section 7.6.8, there is coherence if and only if $P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ satisfy axiom S5(ii), and \underline{E} and $\underline{P}(\cdot|\mathcal{X})$ satisfy C11 and C12, where $\underline{E}(Y) = \underline{P}(P(Y|\Theta))$ is the natural extension. In some cases these axioms can be directly verified. For example, let \underline{P} and $\underline{P}(\cdot|x)$ be the COR models determined by P_0 and $P_0(\cdot|x)$ respectively, with the same value of τ . Provided the linear previsions P_0 , $P(\cdot|\Theta)$ and $P_0(\cdot|\mathcal{X})$ are coherent (e.g. a standard Bayesian model), it can be verified that \underline{P} , $P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ also satisfy the coherence axioms.

The coherence conditions simplify when \mathcal{X} is discrete and $L_x(\theta) = P(\{x\}|\theta)$. Then (by the results in section 8.4) the model is coherent if and only if \underline{E} is \mathcal{X} -conglomerable. Sufficient conditions for coherence are that \mathcal{X} is finite, or that \underline{P} is a lower envelope of countably additive linear previsions and each $P(\cdot|\theta)$ is countably additive.

When \mathcal{X} is continuous and $P(\{x\}|\theta) = 0$ for all x and θ , the coherence axioms reduce to S1 and C11.⁷ One way to establish coherence is to show that \underline{P} is a lower envelope of countably additive priors, and that the sampling models satisfy the regularity conditions in Definition 7.7.1(b). Then, because the posterior defined by the GBR is the lower envelope of Bayesian posteriors, the imprecise model is the lower envelope of standard Bayesian models.

7.8.9 COR-Normal model

Let the sampling models be Normal $(\theta, 1)$, take L_x proportional to the Normal density function, and let P_0 be the Normal (μ, η^2) prior. The Bayesian posterior $P_0(\cdot|x)$ is the Normal $(\alpha(x), \beta^2)$ posterior defined in Example 7.8.3. Let \underline{P} be the COR prior determined by P_0 and τ . By Example 7.8.7, the posteriors $\underline{P}(\cdot|x)$ defined through the GBR are just the COR models

determined by $P_0(\cdot|x)$ and τ . Because \underline{P} is a lower envelope of countably additive priors,⁸ the model $\underline{P}, P(\cdot|\Theta), \underline{P}(\cdot|\mathcal{X})$ is coherent.

Consider the limit of these inferences as the prior variance $\eta^2 \rightarrow \infty$. Then $P_0(\cdot|x)$ tends to the Normal $(x, 1)$ posterior. We know (from section 7.4.4) that these posteriors are not coherent with the Normal sampling model. The following result states that coherence can be achieved by replacing the precise Normal $(x, 1)$ posteriors by COR neighbourhoods.⁹

7.8.10 Coherence of COR-Normal model

Let $P(\cdot|\theta)$ be Normal $(\theta, 1)$. Define the posterior previsions $\underline{P}(\cdot|x)$ to be COR models determined by the Normal $(x, 1)$ posteriors, with constant τ (where $0 < \tau < 1$). Then $P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent.¹⁰

Proof. By Theorem 7.3.6 and results 7.6.8, coherence is equivalent to axiom S1. Suppose this fails, so there is a measurable, non-zero Y such that $P(Y|\theta) > 0$ for all $\theta \in S(Y)$ and $\bar{P}(Y|x) < 0$ for all $x \in T(Y)$. Let f denote the standard Normal density function. Then

$$(a) P(Y|\theta) = \int Y(\theta, x)f(x - \theta)dx > 0 \quad \text{for all } \theta \in S(Y).$$

Let $P_0(\cdot|x)$ be the Normal $(x, 1)$ posterior. From section 2.9.4, we also have

$$(b) P_0(\tau Y^+ + (1 - \tau)Y|x) = \int (\tau Y^+(\theta, x) + (1 - \tau)Y(\theta, x))f(x - \theta)d\theta < 0 \quad \text{for all } x \in T(Y).$$

Use (a) and (b), together with the lemma in section 7.4.4, to show that $\iint Y^+(\theta, x)f(x - \theta)d\theta dx$ is finite. By (a), the same is true when Y^+ is replaced by Y^- or $|Y|$. Now Fubini's theorem can be applied to show that $\int P(Y|\theta)d\theta = \int P_0(Y|x)dx \leq (1 - \tau)^{-1} \int P_0(\tau Y^+ + (1 - \tau)Y|x)dx \leq 0$. If $S(Y)$ has positive Lebesgue measure then by (a) the first expression is positive, a contradiction. If $S(Y)$ has zero Lebesgue measure then

$$P_0(Y|x) = \int_{S(Y)} Y(\theta, x)f(x - \theta)d\theta = 0 \quad \text{for all } x,$$

which is inconsistent with (b). ◆

The coherent posteriors $\underline{P}(\cdot|x)$ are not lower envelopes of standard Bayesian posteriors. Indeed, none of the linear previsions that dominates $\underline{P}(\cdot|x)$ is the posterior for a countably additive prior. To see that, consider the gamble $Y_k(\theta) = A_k(\theta)/f(x - \theta)$, where A_k is the event that $|\theta| \leq k$ and x is fixed. If P is a countably additive prior prevision then $q(x) = \int f(x - \theta)P(d\theta)$ is finite and non-zero, and the standard Bayesian posterior is

$$P(Y_k|x) = q(x)^{-1} \int Y_k(\theta)f(x - \theta)P(d\theta) = q(x)^{-1}P(A_k) \leq q(x)^{-1}.$$

From section 2.9.4,

$$\underline{P}(Y_k|x) \geq (1 - \tau)P_0(Y_k|x) + \tau \inf Y_k = (1 - \tau) \int Y_k(\theta)f(x - \theta)d\theta = 2k(1 - \tau).$$

When k is sufficiently large, $\underline{P}(Y_k|x) > P(Y_k|x)$.

By Theorem 8.2.1, there is a prior prevision \underline{P} such that $\underline{P}, P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent.¹¹ None of the linear previsions in $\mathcal{M}(\underline{P})$ is countably additive. This shows that not all coherent models are lower envelopes of standard Bayesian models.

7.8.11 Imprecise statistical inference

Methods of statistical inference which always yield precise posterior probabilities are unrealistic. A more sensible approach is to construct an imprecise prior and combine this with the sampling model (which may also be imprecise) to produce an imprecise posterior. The initial assessments can be much less precise, and therefore easier to justify than in a Bayesian analysis, but the conclusions may also be weaker. The posterior will usually be more precise than the prior, according to the amount of information in the statistical data.

In applying this method of statistical inference, there are three problems: (a) how should the imprecise prior be constructed?, (b) how should the posterior be constructed?, and (c) are the inferences coherent? One approach to (a) is to form a class \mathcal{M} of linear prior previsions. Another is to directly assess a lower prevision \underline{P} , using the methods in Chapter 4. The second approach may often be easier, because \underline{P} has a behavioural interpretation whereas the priors in \mathcal{M} do not.

Similarly, there are two approaches to problem (b). A posterior class of linear previsions can be constructed by applying Bayes' rule to update the extreme points of $\mathcal{M}(\underline{P})$, or the generalized Bayes rule can be used to directly update \underline{P} . These two approaches are mathematically equivalent.¹²

To determine whether the inferences are coherent, it is necessary in general to check axioms S5 and S7. (These simplify when the sampling models are precise.) The models $\underline{P}, P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent whenever they are lower envelopes of standard Bayesian models, or whenever \mathcal{X} is finite and the posteriors are determined by the generalized Bayes rule. Example 7.8.10 shows that there are coherent statistical models which are not lower envelopes of standard Bayesian models, although it is not yet clear whether any of these models are useful.

7.9 Joint prior previsions

The three previous sections were concerned with a prior prevision \underline{P} which represented beliefs about Θ . We now suppose that \underline{P} is defined on a larger

domain, to represent prior beliefs about both the parameter θ and the future observation x . Again we investigate when \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent.

This problem is particularly important when the sampling model $\underline{P}(\cdot|\Theta)$ is non-linear. In that case prior beliefs about \mathcal{X} are not uniquely determined by the sampling model and prior beliefs about Θ . In other words, a prior prevision \underline{P} on Θ -measurable gambles does not have a unique extension to \mathcal{X} -measurable gambles that is coherent with $\underline{P}(\cdot|\Theta)$. There is then good reason to specify Your prior beliefs about Θ and \mathcal{X} jointly, since these provide more information than Your prior beliefs about Θ alone. When $\underline{P}(\cdot|\Theta)$ is linear, on the other hand, the natural extension $\underline{E}(Y) = \underline{P}(P(Y|\Theta))$ is the unique coherent extension of \underline{P} , and in that case the joint prevision \underline{E} adds no further information to \underline{P} and $\underline{P}(\cdot|\Theta)$.

Formally, we assume that \underline{P} is defined on the same domain \mathcal{F} as $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. Coherence of the triple \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ then reduces to pairwise coherence.

7.9.1 Pairwise coherence theorem

Suppose $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ satisfy assumptions 7.3.2, and \underline{P} is a coherent lower prevision defined on domain \mathcal{F} .¹ Then \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent if and only if they are pairwise coherent,² i.e. $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ satisfy S5, \underline{P} and $\underline{P}(\cdot|\Theta)$ satisfy C11 and C12, \underline{P} and $\underline{P}(\cdot|\mathcal{X})$ satisfy C11 and C12.

Proof. By Theorem 7.1.5, coherence is equivalent to coherence of $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ (S5) plus weak coherence of \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. The weak coherence condition is that

$$\sup U \geq 0 \quad \text{whenever} \quad U = G(X) + G(V|\Theta) + G(W|\mathcal{X}) - G(Z|B), \\ \text{where} \quad B \text{ is } \Omega, \theta \text{ or } x, \text{ and } X, V, W, Z \in \mathcal{F}.$$

Assuming pairwise coherence,

$$\sup U \geq \bar{P}(U) \geq \underline{P}(G(X)) + \underline{P}(G(V|\Theta)) + \underline{P}(G(W|\mathcal{X})) - \underline{P}(G(Z|B)),$$

and $\underline{P}(G(X)) = 0$, $\underline{P}(G(V|\Theta)) \geq 0$ by C11, $\underline{P}(G(W|\mathcal{X})) \geq 0$ by C11, and $\underline{P}(G(Z|B)) = 0$ by C12 or coherence of \underline{P} . Hence $\sup U \geq 0$. Thus pairwise coherence implies coherence. The converse is obvious. ◆

If Θ and \mathcal{X} are both finite, $\underline{P}(\{\theta\}) > 0$ for all θ and $\underline{P}(\{x\}) > 0$ for all x , then (by results 6.5.4 and 7.6.8) coherence reduces to the two versions of C12, the generalized Bayes rule, through which \underline{P} uniquely determines $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$.

When all of P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are linear previsions, coherence is equivalent to axiom S1 plus the axiom

$$(S6') \quad P(Y) = P(P(Y|\Theta)) = P(P(Y|\mathcal{X})) \quad \text{for all } Y \in \mathcal{F}.^3$$

7.9.2 Standard Bayesian models

Suppose that P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are a standard Bayesian model (7.7.1), where P represents beliefs about Θ alone. Define the natural extension to \mathcal{F} by $E(Y) = P(P(Y|\Theta))$. Then the two axioms S1 and S6' (with E replacing P) hold, since these are just the two conditions established in the proof of Theorem 7.7.2. Thus E , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent in a standard Bayesian model.

If E_γ , $P_\gamma(\cdot|\Theta)$ and $P_\gamma(\cdot|\mathcal{X})$ are standard Bayesian models on domain \mathcal{F} for all γ in an index set, by Theorem 7.1.6 their lower envelopes \underline{E} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent. As in section 7.8 we can construct coherent imprecise models as lower envelopes of standard Bayesian models. Here is a simple example.

7.9.3 Imprecise Normal models

Consider the special case of the imprecise Normal model in Example 7.8.5. The sampling models $\underline{P}(\cdot|\theta)$ are lower envelopes of Normal $(\theta + v, 1)$ distributions over $|v| \leq 1$. Your prior beliefs about θ are represented by the lower envelope \underline{P} of $N(\mu, 1)$ distributions with $9 \leq \mu \leq 11$. Your beliefs about the pair (θ, x) are represented by a joint prevision \underline{Q} which is the lower envelope of a class of precise joint distributions. The joint distributions are determined by the marginal $N(\mu, 1)$ distribution for θ and the $N(\theta + v, 1)$ distributions for x conditional on θ , hence by the parameters (μ, v) .

Two different models were considered in Example 7.8.5, and these define two different joint previsions \underline{Q} . In the first ('unlinked') model, \underline{Q}_1 is the lower envelope of all joint distributions (μ, v) with $9 \leq \mu \leq 11$ and $|v| \leq 1$. In the second model, \underline{Q}_2 is the lower envelope of all joint distributions (μ, v) with $\mu = 10 + v$ and $|v| \leq 1$.

The two previsions \underline{Q}_1 and \underline{Q}_2 are quite different. For example, $2\theta - x$ has a precise $N(10, 2)$ marginal distribution under \underline{Q}_2 , but its marginal under \underline{Q}_1 is the lower envelope of $N(\alpha, 2)$ distributions over $8 \leq \alpha \leq 12$. Both \underline{Q}_1 and \underline{Q}_2 have the same Θ -marginal \underline{P} and both are coherent with $\underline{P}(\cdot|\Theta)$. This illustrates that \underline{P} and $\underline{P}(\cdot|\Theta)$ may not have a unique coherent extension to a joint prevision when $\underline{P}(\cdot|\Theta)$ is non-linear.⁴ Specification of the joint prevision \underline{Q} can then provide further information about beliefs.

Two different posteriors $\underline{P}_1(\cdot|\mathcal{X})$ and $\underline{P}_2(\cdot|\mathcal{X})$ were defined in Example 7.8.5. Both models \underline{Q}_i , $\underline{P}(\cdot|\Theta)$ and $\underline{P}_i(\cdot|\mathcal{X})$ (for $i = 1$ or 2) are lower envelopes of standard Bayesian models and therefore coherent.

CHAPTER 8

Statistical reasoning

In statistical inference, we draw conclusions from statistical data about the unknown parameters of a sampling model. Statistical reasoning is a more general process, in which we construct models and make assessments concerning some elements of a statistical problem, and use these to draw conclusions about other elements. In this chapter, we view statistical reasoning as the process of extending a sampling model and other specified previsions, by natural extension, to construct new previsions.

For example, statistical inference is carried out by assessing prior previsions and applying the generalized Bayes rule to construct posterior previsions (sections 8.4, 8.5). This defines a particular strategy for updating beliefs about statistical parameters in the light of observations. It reduces to the familiar Bayesian strategy when both the prior prevision and sampling model are precise. Other types of statistical reasoning can also be viewed as assessment strategies. The two types of extension studied in sections 8.2 and 8.3 define strategies for forming initial beliefs about the parameter or observation.

A general theory of coherent extension is developed in section 8.1. The basic question is: given a coherent collection of conditional lower previsions, what do these imply (through coherence) about other previsions? Provided the specified previsions have some coherent extensions to new previsions, they have minimal coherent extensions that summarize their implications. In the cases of interest, including those already studied in Chapters 3 and 6, the minimal coherent extensions can be constructed directly from the given previsions by natural extension.

The main issues considered in this chapter, for the various types of extension, are as follows:

1. Under what conditions do the specified previsions have coherent extensions?
2. Can the minimal coherent extensions be constructed by natural extension?

3. Can the minimal coherent extensions be constructed as lower envelopes of coherent linear extensions?
4. When are coherent extensions unique, linear or vacuous?

The general theory in section 8.1 settles these problems in the case where all the conditioning partitions are finite. Then coherent extensions always exist, and the minimal coherent extensions can be constructed as natural extensions and as lower envelopes.

Sections 8.2 and 8.3 concern extensions to unconditional (prior) previsions. Coherent extensions always exist. In section 8.2 the prior previsions are constructed from sampling models and posterior previsions. The resulting prior is usually non-vacuous. When the sampling model and posteriors are linear they must be coherent with some linear prior, and their natural extension is the lower envelope of all coherent linear priors. In section 8.3 the sampling model, posteriors and prior prevision for θ are extended to form predictive previsions for x . The posterior previsions are relevant only when the sampling models are non-linear.

Sections 8.4 and 8.5 concern the problem of statistical inference: extending a sampling model and prior to construct posterior previsions. Coherent extension is not always possible and requires some version of \mathcal{X} -conglomerability. Section 8.4 is mainly concerned with the case where the sampling model is linear and prior beliefs are specified about θ alone. These have a unique extension \underline{E} to beliefs about (θ, x) jointly. There is a coherent posterior if and only if \underline{E} is \mathcal{X} -conglomerable, and then the minimal coherent posterior is defined from \underline{E} through the generalized Bayes rule. The minimal coherent posterior depends on the sampling model only through the likelihood function generated by the observation, and it can be written as a lower envelope of Bayesian posteriors defined from linear priors. The results in section 8.5 for non-linear sampling models are similar, provided \underline{E} is taken to be the natural extension of the prior and sampling model and is \mathcal{X} -conglomerable. The resulting posterior depends on the sampling model only through the upper and lower likelihood functions generated by the observation.

The validity of the likelihood principle is examined in section 8.6. When the likelihood function is precise and the observation has positive prior probability, the likelihood principle is a consequence of coherence. We consider justification of the principle for continuous sample spaces, which are regarded as idealizations of discrete spaces. An example is given to show that, unless it is known that the precision of the discrete observation does not depend on the unknown parameter, the continuous versions of the likelihood principle and Bayes' rule are unreliable.

8.1 A general theory of natural extension

The problem considered in this section is to extend a coherent collection of conditional previsions $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ to a larger domain. This includes the problem of extending a model $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_{m-1})$ to previsions conditional on a new partition \mathcal{B}_m , by taking $\underline{P}(\cdot|\mathcal{B}_m)$ to be specified on a trivial domain containing only the zero gamble.

As usual, we interpret the conditional previsions $\underline{P}(\cdot|\mathcal{B}_i)$ as descriptions of behavioural dispositions. We can then ask what these dispositions imply about other dispositions, and attempt to summarize their implications through ‘natural extension’ of each $\underline{P}(\cdot|\mathcal{B}_i)$ to the domain $\mathcal{L}(\Omega)$ containing all gambles. That was done for unconditional previsions in section 3.1, where we saw that a coherent unconditional prevision can always be coherently extended to \mathcal{L} .

We saw in Chapter 6, however, that coherent extension is not always possible when conditional previsions are involved. For example, the ‘uniform distribution’ \underline{P} in Example 6.6.7 is a coherent lower prevision, but there is no conditional prevision $\underline{P}(\cdot|\mathcal{B})$ that is coherent with \underline{P} .

We regard it as a requirement of rationality that specified previsions $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ should have some coherent extensions to the domain \mathcal{L} . The examples in Chapter 6 show that it is not sufficient for this that the specified previsions be coherent. The main tasks of a general theory of coherent extension are (a) to characterize which models do have coherent extensions, and (b) to characterize their minimal coherent extensions. For example, in section 6.8 we saw that \mathcal{B} -conglomerability of a coherent unconditional prevision is necessary and sufficient for coherent extension to a conditional prevision $\underline{P}(\cdot|\mathcal{B})$, and that the minimal coherent extension $\underline{E}(\cdot|\mathcal{B})$ is defined by the generalized Bayes rule.

The results of this section settle the basic issues of extensions for finite partitions \mathcal{B}_i , but otherwise the general theory is much less complete than the theory for unconditional previsions. More complete results can be given for the most important statistical extension problems, in sections 6.7, 6.8, and the rest of this chapter.

Suppose that $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are defined on linear spaces and are coherent. If they have some coherent extensions to larger domains then they have **minimal coherent extensions** which are the lower envelopes of all the coherent extensions. (These are coherent by Theorem 7.1.6.) The minimal coherent extensions are the main object of interest. They may be regarded as models for the behavioural dispositions (prices for contingent gambles) that are effectively implied by the specified previsions.

We wish to construct the minimal coherent extensions from the specified previsions through some sort of ‘natural extension’, as in Definition 3.1.1.

8.1 A GENERAL THEORY OF NATURAL EXTENSION

To do so, consider the justification of the coherence condition 7.1.4. The argument there shows that Your willingness to accept gambles of the form $G(Y_i|\mathcal{B}_i) + Z_i$, where Z_i is non-negative, effectively commits You to pay a specific price α for X contingent on B_0 , where X is an arbitrary gamble and B_0 is in some \mathcal{B}_j . That motivates the following definition of the natural extension $\underline{E}(X|B_0)$ as the largest price α that You are committed to pay through gambles of this form.

8.1.1 Definition of natural extension

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ are defined on linear spaces \mathcal{K}_i for $1 \leq i \leq m$, are separately coherent, and avoid partial loss. For each $X \in \mathcal{L}$ and $B_0 \in \mathcal{B}_j$, define $\underline{E}(X|B_0)$ to be the supremum value of α for which there are Y_i in \mathcal{K}_i such that

$$\sup_{\omega \in B} \left[\sum_{i=1}^m G(Y_i|\mathcal{B}_i) - B_0(X - \alpha) \right] < 0 \quad \text{for all } B \in \bigcup_{i=1}^m S_i(Y_i) \cup \{B_0\}.$$

This defines conditional lower previsions $\underline{E}(\cdot|\mathcal{B}_1), \dots, \underline{E}(\cdot|\mathcal{B}_m)$ on \mathcal{L} , called the **natural extensions** of $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$.¹

The justification in section 7.1.4 shows that, in order to be coherent with the specified previsions $\underline{P}(\cdot|\mathcal{B}_i), \underline{P}(X|B_0)$ must be at least as large as $\underline{E}(X|B_0)$. Any coherent extensions of the specified previsions must therefore dominate the natural extensions. Hence the minimal coherent extensions (if they exist) must dominate the natural extensions. In the statistical problems of interest it turns out that the minimal coherent extensions (when they exist) agree with the natural extensions.

8.1.2 Basic properties of natural extension²

Suppose that $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are defined on linear spaces \mathcal{K}_i , are separately coherent, and avoid partial loss. Let $\underline{E}(\cdot|\mathcal{B}_1), \dots, \underline{E}(\cdot|\mathcal{B}_m)$ be their natural extensions to \mathcal{L} .

- (a) For each $X \in \mathcal{L}$ and $B_0 \in \mathcal{B}_j$, $\inf\{X(\omega): \omega \in B_0\} \leq \underline{E}(X|B_0) \leq \sup\{X(\omega): \omega \in B_0\}$, so that $\underline{E}(\cdot|\mathcal{B}_j)$ is well-defined.
- (b) Each $\underline{E}(\cdot|\mathcal{B}_j)$ is separately coherent.
- (c) Each $\underline{E}(\cdot|\mathcal{B}_j)$ dominates $\underline{P}(\cdot|\mathcal{B}_j)$ on \mathcal{K}_j , i.e. $\underline{E}(X|B_0) \geq \underline{P}(X|B_0)$ when $B_0 \in \mathcal{B}_j$ and $X \in \mathcal{K}_j$.
- (d) $\underline{E}(\cdot|\mathcal{B}_j)$ agrees with $\underline{P}(\cdot|\mathcal{B}_j)$ on \mathcal{K}_j for each $1 \leq j \leq m$ if and only if $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are coherent.
- (e) If $\underline{E}'(\cdot|\mathcal{B}_1), \dots, \underline{E}'(\cdot|\mathcal{B}_m)$ are defined on \mathcal{L} and coherent, and each $\underline{E}'(\cdot|\mathcal{B}_j)$ dominates $\underline{P}(\cdot|\mathcal{B}_j)$ on \mathcal{K}_j , then each $\underline{E}'(\cdot|\mathcal{B}_j)$ dominates $\underline{E}(\cdot|\mathcal{B}_j)$ on \mathcal{L} .

- Hence, if $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ have some coherent extensions to \mathcal{L} , their minimal coherent extensions dominate their natural extensions on \mathcal{L} .
- (f) If $\underline{E}(\cdot|\mathcal{B}_1), \dots, \underline{E}(\cdot|\mathcal{B}_m)$ are coherent then they are the minimal coherent previsions on \mathcal{L} that dominate $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$.

Proof. (a) If $\alpha \geq \sup\{X(\omega): \omega \in B_0\}$ in Definition 8.1.1 then $B_0(X - \alpha) \leq 0$, hence the gambles Y_i incur partial loss (Definition 7.1.2). This shows that $\underline{E}(X|B_0) \leq \sup\{X(\omega): \omega \in B_0\}$. For the other inequality take $\alpha < \inf\{X(\omega): \omega \in B_0\}$, and all $Y_i = 0$ in Definition 8.1.1.

- (b) Verify the axioms of Theorem 6.2.7 for $\underline{E}(\cdot|B_0)$.
 (c) Take $\alpha < \underline{P}(X|B_0)$, $Y_j = B_0 X$, other $Y_i = 0$ in Definition 8.1.1.
 (d) By (c), $\underline{E}(\cdot|\mathcal{B}_j)$ fails to agree with $\underline{P}(\cdot|\mathcal{B}_j)$ on \mathcal{K}_j , just when $\underline{E}(X|B_0) > \underline{P}(X|B_0)$ for some $B_0 \in \mathcal{B}_j, X \in \mathcal{K}_j$. By Definition 8.1.1, that is equivalent to failure of the coherence condition 7.1.4(b).
 (e) If α satisfies the condition in Definition 8.1.1, use $G(Y_i|\mathcal{B}_i) \geq Y_i - \underline{E}'(Y_i|\mathcal{B}_i)$ and coherence of the $\underline{E}'(\cdot|\mathcal{B}_i)$ to give $\alpha < \underline{E}'(X|B_0)$. Hence $\underline{E}(X|B_0) \leq \underline{E}'(X|B_0)$.
 (f) follows from (e). ◆

If the specified previsions are coherent and their natural extensions are coherent, the natural extensions are the minimal coherent extensions to \mathcal{L} . This can fail either because there are no coherent extensions to \mathcal{L} (as in Examples 6.6.6 and 6.6.7), or because there are coherent extensions but the natural extensions are not coherent, as in the following example.

8.1.3 Example

Let Ω be the set of non-zero integers, $B_n = \{n, -n\}$, $\mathcal{B} = \{B_n: n \geq 1\}$. Define a linear prevision P on the linear space containing all gambles with finite support, by $P(Y) = \frac{1}{2}\sum_{n=1}^{\infty} 2^{-n}Y(n)$. Define $\underline{P}(\cdot|\mathcal{B})$ on the trivial space containing only the zero gamble. Let A be the set of positive integers. Since $P(\{n\}) > 0$ and $P(\{-n\}) = 0$, any coherent extensions to all gambles must satisfy the conditions $\underline{P}(\{n\}|B_n) = 1$ for all $n \geq 1$, and hence $\underline{P}(A) = 1$ (using C8). Any extensions $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ which satisfy these conditions are coherent, since they will satisfy axioms C8 and C12. So there is a minimal coherent extension which has each $\underline{P}(\cdot|B_n)$ degenerate on $\{n\}$, and $\underline{P}(A) = 1$.

The natural extension \underline{E} is simply the unconditional natural extension of P , defined in section 3.1. By Corollary 3.4.3, \underline{E} is the lower envelope of all linear extensions of P . There are linear extensions, such as that defined in Example 6.6.6, which assign probability $\frac{1}{2}$ to A , so that $\underline{E}(A) = \frac{1}{2}$. Thus the natural extension \underline{E} is different from the minimal coherent extension \underline{P} . They disagree because \underline{E} reflects only the direct implications of the specified previsions P . In this case P also has implications through coherence for $\underline{P}(\cdot|\mathcal{B})$, which in turn imply that $\underline{P}(A) = 1$. Thus the natural extensions may

8.1 A GENERAL THEORY OF NATURAL EXTENSION

fail to fully reflect the behavioural implications of the specified previsions, and may therefore be incoherent.³

Nevertheless, the natural extensions do coincide with the minimal coherent extensions in some important cases. We show this first for the extension problems already studied in Chapters 3 and 6. Assume that one of the specified previsions is an unconditional prevision \underline{P} , so that an unconditional natural extension \underline{E} is defined. (There is no loss of generality in this assumption since \underline{P} can be specified on a trivial domain.) The following theorem gives a simpler formula for $\underline{E}(\cdot|B_0)$ that holds whenever $\underline{E}(B_0)$ is positive. In that case $\underline{E}(\cdot|B_0)$ is determined by \underline{E} through the generalized Bayes rule.

8.1.4 Theorem

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ satisfy the conditions of Definition 8.1.1 and their natural extensions $\underline{E}(\cdot|\mathcal{B}_i)$ include an unconditional prevision \underline{E} . Then

$$\underline{E}(X) = \sup \left\{ \alpha: X - \alpha \geq \sum_{i=1}^m G(Y_i|\mathcal{B}_i) \text{ for some } Y_i \in \mathcal{K}_i \right\}.$$

If $B_0 \in \mathcal{B}_j, \underline{E}(B_0) > 0$ and $X \in \mathcal{L}$ then

$$\begin{aligned} \underline{E}(X|B_0) &= \sup \left\{ \alpha: B_0(X - \alpha) \geq \sum_{i=1}^m G(Y_i|\mathcal{B}_i) \text{ for some } Y_i \in \mathcal{K}_i \right\} \\ &= \max \{ \beta: \underline{E}(B_0(X - \beta)) \geq 0 \}. \end{aligned}$$

Proof. The formula for $\underline{E}(X)$ follows from 8.1.1 with $B_0 = \Omega$. Applying this formula to $X = B_0$, where $B_0 \in \mathcal{B}_j$ and $\underline{E}(B_0) > 0$, there are $Z_i \in \mathcal{K}_i$ and $\varepsilon > 0$ such that $B_0 - \varepsilon \geq \sum_{i=1}^m G(Z_i|\mathcal{B}_i)$. If $Y_i \in \mathcal{K}_i$ and α satisfy $B_0(X - \alpha) \geq \sum_{i=1}^m G(Y_i|\mathcal{B}_i)$ then, for any $\delta > 0$, $\sum_{i=1}^m G(Y_i + \delta Z_i|\mathcal{B}_i) - B_0(X - \alpha + \delta) \leq \sum_{i=1}^m G(Y_i|\mathcal{B}_i) - B_0(X - \alpha) + \delta \sum_{i=1}^m G(Z_i|\mathcal{B}_i) - \delta B_0 \leq -\delta\varepsilon < 0$. By definition of $\underline{E}(X|B_0)$ this gives $\underline{E}(X|B_0) \geq \alpha - \delta$, hence $\underline{E}(X|B_0) \geq \sup \{ \alpha: B_0(X - \alpha) \geq \sum_{i=1}^m G(Y_i|\mathcal{B}_i) \text{ for some } Y_i \in \mathcal{K}_i \}$. The reverse inequality holds because any α and Y_i satisfying the condition in 8.1.1 must satisfy $\sum_{i=1}^m G(Y_i|\mathcal{B}_i) - B_0(X - \alpha) \leq 0$.

For the last formula, note that \underline{E} is coherent by property 8.1.2(b), hence $\underline{E}(B_0(X - \beta))$ is continuous as a function of β and the maximum is achieved. Suppose that $\underline{E}(B_0(X - \beta)) \geq 0$. For any $\delta > 0$, $\underline{E}(B_0(X - \beta + \delta)) \geq \underline{E}(B_0(X - \beta)) + \delta \underline{E}(B_0) > 0$. Applying the first formula for \underline{E} , $B_0(X - \beta + \delta) \geq \sum_{i=1}^m G(Y_i|\mathcal{B}_i)$ for some $Y_i \in \mathcal{K}_i$, hence $\underline{E}(X|B_0) \geq \beta - \delta$, hence $\underline{E}(X|B_0) \geq \max \{ \beta: \underline{E}(B_0(X - \beta)) \geq 0 \}$. For the reverse inequality, if $B_0(X - \alpha) \geq \sum_{i=1}^m G(Y_i|\mathcal{B}_i)$ then $\underline{E}(B_0(X - \alpha)) \geq 0$ by definition of \underline{E} . ◆

As in section 6.4, when $\underline{E}(B_0)$ is positive there is a unique β such that $\underline{E}(B_0(X - \beta)) = 0$. Thus the natural extension $\underline{E}(\cdot|B_0)$ is uniquely determined

by the unconditional natural extension \underline{E} whenever $\underline{E}(B_0) > 0$, through the generalized Bayes rule.⁴

In the case where only an unconditional prevision \underline{P} is specified on a linear space \mathcal{K} , the natural extension reduces to $\underline{E}(X) = \sup\{\alpha: X - \alpha \geq G(Y), Y \in \mathcal{K}\}$, which agrees with the earlier definition of natural extension (3.1.4).

The next theorem characterizes the natural extensions for the problems studied in Chapter 6.

8.1.5 Theorem

Suppose \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are defined on linear spaces \mathcal{K} and \mathcal{H} , are separately coherent, and avoid sure loss. Their natural extensions are given by the formulas, for all $X \in \mathcal{L}$,

- (a) $\underline{E}(X) = \sup\{\alpha: X - \alpha \geq G(Y) + G(Z|\mathcal{B}), Y \in \mathcal{K}, Z \in \mathcal{H}\}^5$
- (b) when $B_0 \in \mathcal{B}$ and $\underline{E}(B_0) > 0$,

$$\begin{aligned}\underline{E}(X|B_0) &= \sup\{\alpha: B_0(X - \alpha) \geq G(Y) + G(Z|\mathcal{B}), Y \in \mathcal{K}, Z \in \mathcal{H}\} \\ &= \max\{\beta: \underline{E}(B_0(X - \beta)) \geq 0\}\end{aligned}$$

- (c) when $B_0 \in \mathcal{B}$ and $\underline{E}(B_0) = 0$,

$$\underline{E}(X|B_0) = \sup\{\alpha: B_0(X - \alpha) \geq G(Z|B_0), Z \in \mathcal{H}\}.^6$$

As corollaries we obtain the formulas for the minimal coherent extension in the problems of sections 6.7 and 6.8. In both these problems, whenever there are coherent extensions, the minimal coherent extension agrees with the natural extension. Consider first the extension of an unconditional prevision \underline{P} to $\underline{E}(\cdot|\mathcal{B})$, studied in section 6.8.

8.1.6 Corollary

Suppose that \underline{P} is a coherent lower prevision defined on \mathcal{L} and \mathcal{B} is a partition of Ω . Then the natural extension $\underline{E}(\cdot|\mathcal{B})$ of \underline{P} is defined in Theorem 6.8.2(c). (This theorem shows there is some $\underline{P}(\cdot|\mathcal{B})$ coherent with \underline{P} if and only if \underline{P} is \mathcal{B} -conglomerable, and then $\underline{E}(\cdot|\mathcal{B})$ is the minimal conditional prevision coherent with \underline{P} .)

Proof. Define $\underline{P}(\cdot|\mathcal{B})$ on the trivial subspace $\mathcal{H} = \{0\}$. Since \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent their natural extension \underline{E} agrees with \underline{P} . The natural extension $\underline{E}(\cdot|\mathcal{B})$ is given by Theorem 8.1.5. When $\underline{P}(B_0) = 0$, $\underline{E}(X|B_0) = \sup\{\alpha: B_0(X - \alpha) \geq 0\} = \inf\{X(\omega): \omega \in B_0\}$, and this agrees with Theorem 6.8.2.

8.1 A GENERAL THEORY OF NATURAL EXTENSION

8.1.7 Corollary

Suppose that \underline{P} is defined on all \mathcal{B} -measurable gambles and coherent, and $\underline{P}(\cdot|\mathcal{B})$ is defined on \mathcal{L} and separately coherent. Their natural extensions are \underline{E} and $\underline{P}(\cdot|\mathcal{B})$, where $\underline{E}(X) = \underline{P}(\underline{P}(X|\mathcal{B}))$ for all $X \in \mathcal{L}$. (By Theorem 6.7.2, \underline{E} is the minimal extension of \underline{P} that is coherent with $\underline{P}(\cdot|\mathcal{B})$.)

Proof. \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent by Theorem 6.7.2, so that $\underline{E}(\cdot|\mathcal{B})$ agrees with $\underline{P}(\cdot|\mathcal{B})$ on \mathcal{L} by property 8.1.2(d). \underline{E} is given by Theorem 8.1.5(a). If $X - \alpha \geq G(Y) + G(Z|\mathcal{B})$ then $\underline{P}(X|\mathcal{B}) \geq G(Y) + \alpha$ by separate coherence of $\underline{P}(\cdot|\mathcal{B})$, using properties 6.2.6(g, i, j). Hence $\underline{P}(\underline{P}(X|\mathcal{B})) \geq \underline{P}(G(Y)) + \alpha = \alpha$, using coherence of \underline{P} . This proves that $\underline{P}(\underline{P}(X|\mathcal{B})) \geq \underline{E}(X)$. Taking $Z = X$, $Y = \underline{P}(X|\mathcal{B})$ and $\alpha = \underline{P}(\underline{P}(X|\mathcal{B}))$ in Theorem 8.1.5 gives $\underline{E}(X) \geq \underline{P}(\underline{P}(X|\mathcal{B}))$, so there is equality. ◆

The last result shows that an unconditional prevision always has extensions that are coherent with a single conditional prevision. It generalizes as follows.

8.1.8 Unconditional extension theorem

Suppose that $\underline{P}, \underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are defined on linear spaces \mathcal{K}_i and are coherent. Let \underline{E} be their natural extension as an unconditional lower prevision to \mathcal{L} , given by Theorem 8.1.4. Then $\underline{E}, \underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are coherent, so that \underline{E} is the minimal extension of \underline{P} to \mathcal{L} that is coherent with $\underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$.⁷

Proof. By Theorem 7.1.5 it is enough to prove that $\underline{E}, \underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are weakly coherent. Suppose not. Then there are $Y_1 \in \mathcal{L}$, $Y_i \in \mathcal{K}_i$ for $2 \leq i \leq m$, B_0, Y_0 and $\delta > 0$ such that $Y_1 - \underline{E}(Y_1) + \sum_{i=2}^m G(Y_i|\mathcal{B}_i) - G(Y_0|B_0) + 2\delta \leq 0$. Use Theorem 8.1.4 to approximate $\underline{E}(Y_1)$ by gambles $Z_i \in \mathcal{K}_i$, such that $Y_1 - \underline{E}(Y_1) \geq G(Z_1) + \sum_{i=2}^m G(Z_i|\mathcal{B}_i) - \delta$. Using separate coherence of the $\underline{P}(\cdot|\mathcal{B}_i)$, $G(Z_1) + \sum_{i=2}^m G(Y_i + Z_i|\mathcal{B}_i) - G(Y_0|B_0) + \delta \leq 0$.

If $B_0 \in \mathcal{B}_j$ and $Y_0 \in \mathcal{K}_j$ for some $j \geq 2$, this contradicts weak coherence of $\underline{P}, \underline{P}(\cdot|\mathcal{B}_2), \dots, \underline{P}(\cdot|\mathcal{B}_m)$. If $B_0 = \Omega$ and $Y_0 \in \mathcal{L}$ then $Y_0 - \underline{E}(Y_0) - \delta \geq G(Z_1) + \sum_{i=2}^m G(Y_i + Z_i|\mathcal{B}_i)$, which contradicts Theorem 8.1.4. ◆

For example, if coherent sampling models $\underline{P}(\cdot|\Theta)$ and posterior previsions $\underline{P}(\cdot|\mathcal{X})$ are specified, You can always construct an unconditional prevision \underline{E} on $\mathcal{L}(\Theta \times \mathcal{X})$ that is coherent with them, even if Θ and \mathcal{X} are infinite spaces. More generally, You might assess many conditional previsions $\underline{P}(\cdot|\mathcal{B}_i)$ in order to construct unconditional previsions. You can always construct a coherent extension \underline{E} , even if \mathcal{B}_i are infinite partitions. The reason for this is that only previsions conditional on a finite partition are extended. (We know from Chapter 6 that coherent extension to $\underline{P}(\cdot|\mathcal{B})$ may

not be possible when \mathcal{B} is infinite.) The next theorem states that all the specified previsions can be coherently extended provided all the partitions are finite.

8.1.9 Finite extension theorem

If all the partitions $\mathcal{B}_1, \dots, \mathcal{B}_m$ are finite and $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are coherent, then their natural extensions $\underline{E}(\cdot|\mathcal{B}_1), \dots, \underline{E}(\cdot|\mathcal{B}_m)$ are coherent and are the minimal coherent extensions.⁸

The finite extension theorem settles the problem of coherent extension when all the partitions involved are finite. (For example, in statistical problems where both Θ and \mathcal{X} are finite, or more generally whenever the overall possibility space Ω is finite.)

Next consider when the minimal coherent extensions can be constructed as the lower envelopes of coherent collections of linear previsions. Theorem 3.4.1 asserts that the natural extension of an unconditional prevision P is the lower envelope of the class $\mathcal{M}(P)$ of dominating linear previsions. That is often useful in constructing the natural extension. Using Williams' theorem (Appendix K3) we can show that Theorem 3.4.1 generalizes to the case where all the partitions \mathcal{B}_i are finite.⁹ In that case the natural extensions $\underline{E}(\cdot|\mathcal{B}_1), \dots, \underline{E}(\cdot|\mathcal{B}_m)$ are the lower envelopes of all coherent collections of linear previsions $P(\cdot|\mathcal{B}_1), \dots, P(\cdot|\mathcal{B}_m)$ that dominate $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$.

8.1.10 Lower envelope theorem

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ are defined on linear spaces \mathcal{K}_i for $1 \leq i \leq m$ and are coherent, where all the partitions \mathcal{B}_i are finite. Let Γ index the class of all coherent collections of linear previsions $P_\gamma(\cdot|\mathcal{B}_1), \dots, P_\gamma(\cdot|\mathcal{B}_m)$, each defined on \mathcal{L} , such that $P_\gamma(X|B) \geq \underline{P}(X|B)$ for all $X \in \mathcal{K}_i$, $B \in \mathcal{B}_i$ and $1 \leq i \leq m$. Then the natural extensions (which are the minimal coherent extensions) are lower envelopes over the class Γ , i.e. $\underline{E}(X|B) = \inf \{P_\gamma(X|B) : \gamma \in \Gamma\}$ for all $X \in \mathcal{L}$, $B \in \mathcal{B}_i$ and $1 \leq i \leq m$.¹⁰

Proof. By Appendix K3, the lower envelopes of Γ are coherent extensions of the given previsions to \mathcal{L} . They are the minimal coherent extensions because, again using K3, any coherent extensions are the lower envelopes of some class of coherent linear collections, and this must be a subclass of Γ . By Theorem 8.1.9, the natural extensions are the minimal coherent extensions. ◆

In the setting of Chapter 6, for example, suppose that \mathcal{B} is finite and \underline{P} , $P(\cdot|\mathcal{B})$ are coherent. Then their minimal coherent extensions to \mathcal{L} are the

8.2 EXTENSION TO PRIOR PREVISIONS

natural extensions $\underline{E}, \underline{E}(\cdot|\mathcal{B})$ defined in Theorem 8.1.5, and these are the lower envelopes of all coherent linear pairs $P, P(\cdot|\mathcal{B})$ such that P dominates \underline{P} and $P(\cdot|\mathcal{B})$ dominates $\underline{P}(\cdot|\mathcal{B})$.

When some of the partitions \mathcal{B}_i are infinite, the last two theorem can fail in various ways:

1. The specified previsions may have no coherent extensions (e.g. 6.6.6 and 6.6.7).
2. The specified previsions may not be lower envelopes of coherent linear collections (6.6.9, 6.6.10).
3. The minimal coherent extensions may differ from the natural extensions (8.1.3).
4. The minimal coherent extensions may not be lower envelopes of coherent linear collections.¹¹

It is an open problem, for general partitions \mathcal{B}_i , to characterize the minimal coherent extensions of specified previsions $\underline{P}(\cdot|\mathcal{B}_i)$, and to find necessary and sufficient conditions for (1) the existence of coherent extensions, (2) coherence of the natural extensions, (3) the minimal coherent extensions to be lower envelopes of coherent linear collections. In the following sections we consider these problems for the most important types of statistical extension.

8.2 Extension to prior previsions

In this and the following sections we examine specific kinds of extension that are important in statistical problems. Two kinds of extension are covered by the results of Chapter 6:

1. The extension of sampling models and prior beliefs about θ to prior beliefs about (θ, x) jointly.¹
2. The extension of prior beliefs about (θ, x) to posterior beliefs about θ after observing x .²

In this section we consider natural extension of specified sampling models and posterior previsions to construct unconditional (prior) previsions. That is, we ask: what do the specified sampling models and posteriors imply about prior beliefs? Such extensions are useful primarily in assessment strategies for constructing beliefs about θ or about a future observation x .³

This extension problem is of theoretical interest also because the objective statistical methods discussed in section 7.4 generate precise posterior previsions from sampling models plus prior ignorance. We saw in section 7.4 that the sampling models and posteriors are often incoherent, but when they are coherent they have a coherent natural extension to prior beliefs

about Θ . It is important to examine whether these are realistic assessments of prior evidence. The prior previsions generated by objective methods are typically non-vacuous and therefore inconsistent with complete ignorance, and they may be highly unrealistic.

The basic result is that coherent sampling models $\underline{P}(\cdot|\Theta)$ and posteriors $\underline{P}(\cdot|\mathcal{X})$ always have a coherent extension to an unconditional prevision, and their natural extension \underline{E} is the minimal coherent extension. Thus \underline{E} summarizes the prices for unconditional gambles that are implied by $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. That is formally stated in the next theorem, which also gives general formulas for \underline{E} in terms of $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$.

8.2.1 Prior extension theorem

Suppose that $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ satisfy assumptions 7.3.2 and are coherent, so each is defined on a linear space \mathcal{F} of measurable gambles and they satisfy axiom S5. Let \underline{E} be their natural extension to an unconditional prevision on $\mathcal{L} = \mathcal{L}(\Theta \times \mathcal{X})$. Then \underline{E} is the minimal lower prevision on \mathcal{L} that is coherent with $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. For gambles X in \mathcal{F} , \underline{E} is given by the formulas

$$\begin{aligned}\underline{E}(X) &= \sup \{\inf \underline{P}(Y|\Theta): Y \in \mathcal{F}, \underline{P}(X - Y|\mathcal{X}) \geq 0\} \\ &= \sup \{\inf \underline{P}(Z|\mathcal{X}): Z \in \mathcal{F}, \underline{P}(X - Z|\Theta) \geq 0\}.\end{aligned}$$

Proof. \underline{E} is the minimal coherent extension by the unconditional extension theorem 8.1.8. By Theorem 8.1.4,

$$\underline{E}(X) = \sup \{\alpha: X - \alpha \geq G(U|\Theta) + G(V|\mathcal{X}) \text{ for some } U, V \in \mathcal{F}\}.$$

For such α , U and V , define $Y = X - G(V|\mathcal{X})$. Then $Y \in \mathcal{F}$ by 7.3.2, and $\underline{P}(Y|\Theta) \geq \alpha$. Also $\underline{P}(X - Y|\mathcal{X}) = \underline{P}(G(V|\mathcal{X})|\mathcal{X}) = 0$. This shows that $\underline{E}(X) \leq \sup \{\inf \underline{P}(Y|\Theta): Y \in \mathcal{F}, \underline{P}(X - Y|\mathcal{X}) \geq 0\}$.

For the reverse inequality, suppose that $Y \in \mathcal{F}$ and $\underline{P}(X - Y|\mathcal{X}) \geq 0$. Let $U = Y$, $V = X - Y$, $\alpha = \inf \underline{P}(Y|\Theta)$. Then $G(U|\Theta) + G(V|\mathcal{X}) = Y - \underline{P}(Y|\Theta) + X - Y - \underline{P}(X - Y|\mathcal{X}) \leq X - \alpha$, so that $\underline{E}(X) \geq \alpha = \inf \underline{P}(Y|\Theta)$. This gives the first formula. The second formula follows by reversing the roles of Θ and \mathcal{X} . ♦

The rationale behind the first formula is as follows. If $\underline{P}(X - Y|\mathcal{X}) \geq 0$ then You must almost prefer X to Y because $\underline{P}(X - Y) \geq \inf \underline{P}(X - Y|\mathcal{X}) \geq 0$ for any coherent extension \underline{P} . Then $\underline{P}(X) \geq \underline{P}(X - Y) + \underline{P}(Y) \geq \underline{P}(Y) \geq \inf \underline{P}(Y|\Theta)$ by coherence, so that You must be willing to pay up to $\inf \underline{P}(Y|\Theta)$ for X . The natural extension $\underline{E}(X)$ is just the supremum of such prices.

The natural extension to prior beliefs about Θ is of particular interest. This is just the restriction of \underline{E} to the linear space \mathcal{K} of \mathcal{C} -measurable

8.2 EXTENSION TO PRIOR PREVISIONS

gambles. For X in \mathcal{K} the second formula in the theorem simplifies slightly to

$$\underline{E}(X) = \sup \{\inf \underline{P}(Z|\mathcal{X}): Z \in \mathcal{F}, X \geq \bar{P}(Z|\Theta)\}.$$

8.2.2 Prior ignorance

When are the sampling models and posteriors coherent with vacuous prior beliefs about Θ ? That happens just when their natural extension \underline{E} is vacuous on \mathcal{K} . For example, if the posteriors are vacuous then \underline{E} is vacuous on \mathcal{K} , by section 7.6.7. That is not typical, however. Since \underline{E} is coherent with $\underline{P}(\cdot|\mathcal{X})$, axiom C8 implies $\underline{E}(Y) \leq \sup \bar{P}(Y|\mathcal{X})$ for all $Y \in \mathcal{K}$. By taking $Y = \{\theta\}$, a necessary condition for \underline{E} to be vacuous is that $\sup \bar{P}(\{\theta\}|\mathcal{X}) = 1$ for every $\theta \in \Theta$.⁴ Thus every possible parameter value must have posterior probabilities arbitrarily close to one. That can happen only when the posteriors are close to vacuous or highly degenerate. Otherwise, the natural extension \underline{E} is nonvacuous on \mathcal{K} , and the sampling models and posteriors are inconsistent with ‘prior ignorance’ about Θ . Objective statistical methods that assume prior ignorance actually imply non-vacuous prior beliefs. The following is a striking example of this.

8.2.3 Binomial model with improper prior (Example 7.4.8)

Let $P(\cdot|\theta)$ be the binomial (n, θ) sampling models and $P(\cdot|x)$ the posteriors for the improper prior density $\theta^{-1}(1-\theta)^{-1}$, defined for $1 \leq x \leq n-1$. Assuming that the posteriors for $x=0$ and $x=n$ are defined as (coherent) linear previsions,⁵ Theorem 8.2.5 implies that the natural extension \underline{E} is the lower envelope of all linear previsions P on \mathcal{F} that are coherent with $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$. We show that any such P must assign probability one to the event that θ is arbitrarily close to zero or one, so that \underline{E} assigns lower probability one to this event.

For fixed x ($1 \leq x \leq n-1$) and $\delta > 0$, define $L(\theta) = P(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

and $A(\delta) = \{\theta: \delta \leq \theta \leq 1-\delta\}$. Then $Y(\delta) = A(\delta)L^{-1}$ is a gamble in \mathcal{K} . The posterior densities are $g(\theta|x) = q(x)^{-1}\theta^{-1}(1-\theta)^{-1}L(\theta)$, so that $P(Y(\delta)|x) = q(x)^{-1} \int_{\delta}^{1-\delta} \theta^{-1}(1-\theta)^{-1} d\theta$. Hence $P(Y(\delta)|x) \rightarrow \infty$ as $\delta \rightarrow 0$. Using coherence, $P(\{x\})P(Y(\delta)|x) = P(\{x\}Y(\delta)) = P(Y(\delta)P(\{x\}|\Theta)) = P(Y(\delta)L) = P(A(\delta)) \leq 1$. Hence $P(\{x\}) = 0$ and $P(A(\delta)) = 0$.

Thus any coherent linear prior P must assign zero probability to the event that $1 \leq x \leq n-1$, and also to the events $A(\delta)$ that $\delta \leq \theta \leq 1-\delta$, for every positive δ . If P is countably additive it follows that P assigns probability one to the hypothesis that the sampling process is deterministic (θ is zero or one).⁶ In fact, if the linear prior P is not completely degenerate, the

postiors for $x=0$ and $x=n$ must be essentially degenerate at $\theta=0$ and $\theta=1$ respectively, as assumed in Example 7.4.8.⁷

The inferences considered in Example 7.4.8, based on the improper prior, will therefore be unreasonable in almost all practical contexts, despite their coherence. They effectively imply that You are initially sure that the chance θ is either zero or one, that You will not observe both successes and failures in the Bernoulli trials, and hence that You will not actually use the improper prior to generate a posterior! The inferences will be inconsistent with prior evidence ('externally irrational'), unless this shows that the experiment actually is deterministic.⁸

8.2.4 Linear priors

When can $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ be coherently extended to a linear prior? For discrete sample spaces, extension to a linear prior is typically impossible unless the posteriors $\underline{P}(\cdot|\mathcal{X})$ are linear. In fact, if there is $x \in \mathcal{X}$ such that the posterior $\underline{P}(\cdot|x)$ is non-linear and $\inf\{\underline{P}(\{x\}|\theta) : \theta \in \Theta\} > 0$, then any coherent prior \underline{P} on \mathcal{F} must be non-linear.⁹ Hence, if \mathcal{X} is discrete, the sampling model $P(\cdot|\Theta)$ is linear and some posterior $\underline{P}(\cdot|x)$ is non-linear, any coherent prior will be non-linear on \mathcal{X} . (Otherwise it would have a linear extension to \mathcal{F} .) In that case, imprecise posterior beliefs about Θ are inconsistent with precise prior beliefs about Θ .¹⁰

Suppose then that both $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are linear previsions on \mathcal{F} . Their natural extension \underline{E} need not be linear, but the next result shows that it is the lower envelope of all linear priors that are coherent with $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$.

8.2.5 Lower envelope theorem

Suppose that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are linear previsions satisfying assumptions 7.3.2 and are coherent. Let \underline{E} be their natural extension to \mathcal{L} . Then a linear prevision P on \mathcal{X} is coherent with $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ if and only if P dominates \underline{E} on \mathcal{X} . The class of all linear previsions on \mathcal{L} that are coherent with $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ is just $\mathcal{M}(\underline{E})$.

Proof. Since \underline{E} , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent they satisfy S7(i) of Theorem 7.6.5, so that $\bar{E}(P(Y|\Theta)) \leq \sup P(Y|\mathcal{X})$ for all $Y \in \mathcal{F}$. When P dominates \underline{E} on \mathcal{X} we have $P(P(Y|\Theta)) \leq \sup P(Y|\mathcal{X})$ for all $Y \in \mathcal{F}$, so P , $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent by Corollary 7.6.3. Conversely, if P is coherent with $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ then P can be coherently extended to \mathcal{L} by the unconditional extension theorem, and P must dominate \underline{E} by minimality of \underline{E} . The second assertion can be proved using 7.1.4 and 7.1.5. ◆

8.2 EXTENSION TO PRIOR PREVISIONS

Because \underline{E} is coherent, $\mathcal{M}(\underline{E})$ is non-empty. The theorem therefore tells us that coherent linear previsions $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are always coherent with some linear prior prevision.¹¹ This formalizes an idea found in the Bayesian literature, that assessments of posterior probabilities should be consistent with some prior distribution.¹² However, it is important to make the following qualifications:

1. Linear previsions $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ need not determine a unique linear prior, so their natural extension, which is the lower envelope of all linear extensions, may be non-linear (8.2.6).
2. There may be no countably additive linear prior.¹³
3. Even if the prior and sampling models are countably additive, they may not determine $P(\cdot|\mathcal{X})$ as the unique coherent posterior (see sections 7.7.4 and 8.6.5). Statistical inferences involving linear previsions need not be 'Bayesian' in the sense that they proceed via Bayes' rule.
4. When \mathcal{X} is discrete and some posterior $\underline{P}(\cdot|x)$ is non-linear, there will typically be no coherent linear prior (8.2.4).

8.2.6 Unique linear prior

Suppose $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent linear previsions. When do they determine a unique linear prior prevision? Equivalently (by 8.2.5), when is their natural extension \underline{E} linear?

To see that \underline{E} is not always linear, let $\Theta = \mathcal{X}$ be an arbitrary space, and take all the conditional previsions to be degenerate, $P(\{\theta\}|\theta) = P(\{x\}|x) = 1$. These posteriors are intuitively reasonable for any prior, and it is easy to verify that any prior is coherent with $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$. The natural extension \underline{E} is vacuous.

The following result shows that there is often a unique linear prior when the sample space is discrete.

8.2.7 Uniqueness theorem

Suppose that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent linear previsions, and there is x such that the likelihood function $L(\theta) = P(\{x\}|\theta)$ is bounded away from zero. Then the natural extension \underline{E} is linear on \mathcal{F} and is the unique coherent linear prior on \mathcal{F} . It is determined on \mathcal{X} by $P(\cdot|x)$ and L , as $E(Z) = P(ZL^{-1}|x)/P(L^{-1}|x)$ for $Z \in \mathcal{X}$.

Proof. For any coherent linear prior P , $P(YL) = P(Y|x)P(L)$ for all $Y \in \mathcal{X}$, by applying S6* to $\{x\}Y$. Since L is bounded away from zero, L^{-1} is in \mathcal{X} .

Given any $Z \in \mathcal{K}$, take $Y = ZL^{-1}$ to give $P(Z) = P(ZL^{-1}|x)P(L)$. The choice $Z = 1$ gives $P(L) = 1/P(L^{-1}|x)$. Thus P is uniquely determined on \mathcal{K} , and hence also on \mathcal{F} by $P(Z) = P(P(Z|\Theta))$, and P is linear on \mathcal{F} . By Theorem 8.2.5, $\underline{E} = P$ on \mathcal{F} . ♦

If also Θ is finite, the unique linear prior can be defined by $E(\{\theta\}) = c(x)P(\{\theta\}|x)/P(\{x\}|\theta)$ for all $\theta \in \Theta$, where $c(x)^{-1} = \sum_{\psi \in \Theta} P(\{\psi\}|x)/P(\{x\}|\psi)$. (Take $Z = \{\theta\}$ in the theorem.)

By reversing the roles of Θ and \mathcal{X} in the theorem, there is a unique linear prior on \mathcal{F} provided there is some θ whose posterior probabilities are bounded away from zero. That will often hold when Θ is a discrete space. Thus the prior will often be uniquely determined by $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ provided either Θ or \mathcal{X} is discrete.¹⁴

8.2.8 Beta-binomial model

As an example, consider the binomial (n, θ) sampling models and beta (s', t') posteriors defined in Example 7.7.3. Here the likelihood functions

$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$ are not bounded away from zero, so the last theorem cannot be applied directly. But the likelihood functions are bounded away from zero on any closed subset A of $(0, 1)$. Take $Y = AL^{-1}$ in the proof of Theorem 8.2.7 to give $P(A) = P(AL^{-1}|x)P(L) = c(x) \int_A \theta^{s'-1} (1-\theta)^{t'-1} d\theta$. Because this holds for all closed subsets A ,¹⁵ any coherent linear prior P has density proportional to $\theta^{s'-1} (1-\theta)^{t'-1}$, which is the beta (s, t) density. Thus the binomial sampling models and beta posteriors determine a unique linear prior, the beta (s, t) prior.

8.2.9 Normal distributions

Consider the inferences from the improper uniform prior that were criticized in section 7.4.4. The sampling distributions are Normal $(\theta, 1)$, and the posterior distributions are Normal $(x, 1)$. These models are not coherent, but they are weakly coherent, and so they have a natural extension to prior previsions \underline{E} that are weakly coherent with them.¹⁶ The natural extension is defined for Borel-measurable gambles X by $\underline{E}(X) = \lim_{c \rightarrow \infty} \inf_{u \in \mathbb{R}} (2c)^{-1} \int_{u-c}^{u+c} X(\theta) d\theta$.¹⁷

Proof. Let α denote the last expression. We need to show $\underline{E}(X) = \alpha$. By the second formula in Theorem 8.2.1, $\underline{E}(X) = \sup \{\inf P(Z|\mathcal{X}): Z \in \mathcal{F}, P(Z|\Theta) \leq X\}$. Suppose $P(Z|\Theta) \leq X$ and $\inf P(Z|\mathcal{X}) = \beta$. Consider $\zeta(c, u) = \int_{u-c}^{u+c} P(Z|\theta) d\theta - \int_{u-c}^{u+c} P(Z|x) dx$. By a simple extension of the lemma in section 7.4.4, there is a constant K such that $|\zeta(c, u)| \leq K$ for all positive c and real u . Hence

8.3 EXTENSION TO PREDICTIVE PREVISIONS

$$\begin{aligned} \beta &\leq (2c)^{-1} \int_{u-c}^{u+c} P(Z|x) dx = (2c)^{-1} \left[\int_{u-c}^{u+c} P(Z|\theta) d\theta - \zeta(c, u) \right] \\ &\leq (2c)^{-1} \left[\int_{u-c}^{u+c} X(\theta) d\theta + K \right]. \end{aligned}$$

By minimizing the last expression over u , and taking limits as $c \rightarrow \infty$, we see that $\alpha \geq \beta$. This shows that $\alpha \geq \underline{E}(X)$.

For the reverse inequality, let

$$Z(\theta, x) = \{(\theta, x): |x - \theta| \leq c\} (2c)^{-1} (f(x - \theta))^{-1} X(\theta),$$

where f is the standard Normal density function. Then Z is a bounded measurable function, with

$$\begin{aligned} P(Z|\theta) &= \int_{-\infty}^{\infty} Z(\theta, x) f(x - \theta) dx = (2c)^{-1} X(\theta) \int_{\theta-c}^{\theta+c} dx = X(\theta), \\ P(Z|x) &= \int_{-\infty}^{\infty} Z(\theta, x) f(x - \theta) d\theta = (2c)^{-1} \int_{x-c}^{x+c} X(\theta) d\theta. \end{aligned}$$

Applying the formula from Theorem 8.2.1, $\underline{E}(X) \geq \inf P(Z|\mathcal{X}) = (2c)^{-1} \inf_{x \in \mathbb{R}} \int_{x-c}^{x+c} X(\theta) d\theta$. Taking limits as $c \rightarrow \infty$, $\underline{E}(X) \geq \alpha$. ♦

The prior prevision \underline{E} can be regarded as a proper replacement for the improper uniform density that generated the Normal posteriors. Clearly \underline{E} is translation-invariant. It is a kind of ‘uniform distribution’ on the real line (section 2.9.7).

The linear prior previsions that are weakly coherent with the Normal sampling models and posteriors are just those which dominate \underline{E} .¹⁸ Loosely, these are the linear previsions that can be obtained as limits of uniform distributions on increasingly long finite intervals.¹⁹ None of the priors P in $\mathcal{M}(\underline{E})$ is countably additive, because $P(A) = \underline{E}(A) = \bar{E}(A) = 0$ for all sets A with finite Lebesgue measure. All these linear priors are translation-invariant, but (as noted in section 3.5.8) the class of translation-invariant linear previsions is much larger than $\mathcal{M}(\underline{E})$.

The properties of \underline{E} and its dominating linear previsions indicate that they are not reasonable models for prior beliefs about a real parameter. (They assign upper probability zero to every bounded set, and they are not countably additive or fully conglomerable.) This confirms that the inferences from the improper uniform prior are unreasonable.

8.3 Extension to predictive previsions

Next we consider extension of a sampling model, posterior beliefs and prior beliefs about θ to construct prior beliefs about θ and x jointly.¹ That is, in

addition to the assessments $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ assumed in section 8.2, we suppose that prior beliefs about θ are assessed in the form of a prior prevision \underline{P} defined on the linear space \mathcal{K} . We assume that the specified previsions \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent, i.e. they satisfy axioms S5 and S7. The aim is to define a coherent extension of \underline{P} to \mathcal{F} , representing prior beliefs about (θ, x) jointly.

Natural extensions of this type are useful primarily for constructing beliefs about the future observation x from the other assessments. These beliefs are represented by the \mathcal{X} -marginal of the natural extension \underline{E} , which is called the **predictive prevision** for x .

The extension problem is straightforward when the sampling models are linear. Then \underline{P} and $\underline{P}(\cdot|\Theta)$ have a unique coherent extension to \mathcal{F} defined by $\underline{E}(Y) = \underline{P}(P(Y|\Theta))$, and the posteriors $\underline{P}(\cdot|\mathcal{X})$ cannot add any further information to constrain \underline{E} . We are therefore concerned mainly with the case of non-linear sampling models, for which the posteriors may add further information to sharpen the predictive prevision.

The basic result is that coherent extensions of this type always exist, and the natural extension is the minimal coherent extension.

8.3.1 Predictive extension theorem²

Suppose that \underline{P} is specified on \mathcal{K} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are specified on \mathcal{F} , assumptions 7.3.2 and 7.6.1 are satisfied, and the given previsions are coherent. Let \underline{E} be their natural extension to an unconditional prevision on \mathcal{L} . Then \underline{E} is the minimal extension of \underline{P} that is coherent with $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. For gambles X in \mathcal{F} , \underline{E} is given by the formulas $\underline{E}(X) = \sup \{\underline{P}(P(Y|\Theta)): Y \in \mathcal{F}, \underline{P}(X - Y|\mathcal{X}) \geq 0\}$ and $\underline{E}(X) = \sup \{\inf \underline{P}(Z|\mathcal{X}): Z \in \mathcal{F}, \underline{P}(P(X - Z|\Theta)) \geq 0\}$. The predictive prevision is defined for all \mathcal{S} -measurable gambles X by $\underline{E}(X) = \sup \{\underline{P}(P(Y|\Theta)): Y \in \mathcal{F}, X \geq \bar{P}(Y|\mathcal{X})\}$.

The rationale behind the first formula for \underline{E} is that when $\underline{P}(X - Y|\mathcal{X}) \geq 0$, X must be almost preferred to Y . Any coherent extension \underline{E} must then satisfy $\underline{E}(X - Y) \geq 0$, hence $\underline{E}(X) \geq \underline{E}(X - Y) + \underline{E}(Y) \geq \underline{E}(Y) \geq \underline{P}(P(Y|\Theta))$. The minimal coherent extension achieves this lower bound.

Compare the formulas for \underline{E} with those in Theorem 8.2.1. We have $\underline{P}(P(Y|\Theta)) \geq \inf \underline{P}(Y|\Theta)$ so that (as expected) the natural extension of \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ is at least as precise as the natural extension of $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$.

Note that \underline{E} depends on \underline{P} and $\underline{P}(\cdot|\Theta)$ only through their natural extension to \mathcal{F} , defined by $\underline{Q}(Y) = \underline{P}(P(Y|\Theta))$. The next result shows that \underline{E} dominates \underline{Q} , and \underline{E} agrees with \underline{Q} when $P(\cdot|\Theta)$ is linear.

8.3.2 Lemma

Under the assumptions of Theorem 8.3.1, \underline{E} satisfies the inequalities $\underline{P}(\underline{P}(X|\Theta)) \leq \underline{E}(X) \leq \underline{P}(\bar{P}(X|\Theta))$ for all X in \mathcal{F} .³

Proof. For the first inequality take $Y = X$ in the first formula of Theorem 8.3.1. For the second inequality suppose $Y \in \mathcal{F}$ and $\underline{P}(X - Y|\mathcal{X}) \geq 0$. Then $\underline{P}(\bar{P}(X - Y|\Theta)) \geq 0$ by S7(i) of Theorem 7.6.5. Hence $\underline{P}(\bar{P}(X|\Theta)) \geq \underline{P}(\bar{P}(X - Y|\Theta) + \underline{P}(Y|\Theta)) \geq \underline{P}(\bar{P}(X - Y|\Theta)) + \underline{P}(P(Y|\Theta)) \geq \underline{P}(P(Y|\Theta))$. By the first formula in 8.3.1, $\underline{P}(\bar{P}(X|\Theta)) \geq \underline{E}(X)$. ◆

If the sampling model $P(\cdot|\Theta)$ is linear then $\underline{E}(X) = \underline{P}(P(X|\Theta))$ defines the unique extension of \underline{P} to \mathcal{F} that is coherent with $P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. In that case, by 8.3.1, the natural extension of the prior and linear sampling model is automatically coherent with the posteriors.

The following example illustrates the case of linear sampling models.

8.3.3 Normal model (Example 7.8.3)

Suppose that the sampling distributions are $N(\theta, 1)$, the prior distribution for θ is $N(\mu, \eta^2)$, and the posterior distributions are $N(\alpha(x), \beta^2)$ where $\beta^{-2} = \eta^{-2} + 1$, $\alpha(x) = \mu + (x - \mu)\beta^2$. These determine a unique coherent predictive distribution for x , which is $N(\mu, \eta^2 + 1)$. If we replace the prior distribution by an imprecise prior \underline{P} which is the lower envelope of $N(\mu, \eta_\gamma^2)$ distributions over $\gamma \in \Gamma$, the predictive prevision for x is still uniquely determined, by $\underline{E}(X) = \underline{P}(P(X|\Theta)) = \inf \{\underline{P}_\gamma(P(X|\Theta)): \gamma \in \Gamma\}$. This is just the lower envelope of $N(\mu, \eta_\gamma^2 + 1)$ distributions over $\gamma \in \Gamma$.

When the sampling models $P(\cdot|\Theta)$ are non-linear the posteriors may not be uniquely determined by \underline{P} and $\underline{P}(\cdot|\Theta)$, and they can then provide extra information which sharpens the natural extension. That is illustrated by the next example.

8.3.4 Imprecise Normal model (Examples 7.8.5, 7.9.3)

Consider the second model defined in Example 7.8.5. The prior \underline{P} is the lower envelope of $N(\mu, 1)$ distributions over $9 \leq \mu \leq 11$, the sampling models $P(\cdot|\Theta)$ are lower envelopes of $N(\theta + v, 1)$ over $|v| \leq 1$, and the linear posteriors $P(\cdot|x)$ are $N(5 + \frac{1}{2}x, \frac{1}{2})$. These are coherent by 7.8.5. Let \underline{E} be their natural extension. The posterior distribution of $2\theta - x$ conditional on each x is $N(10, 2)$. For coherence, $2\theta - x$ must have the prior $N(10, 2)$ distribution under \underline{E} .⁴ If A is the event that $2\theta - x \geq 10$, for instance, we obtain the precise prior probabilities $\underline{E}(A) = \bar{E}(A) = \frac{1}{2}$.

To see that the natural extension of \underline{P} and $P(\cdot|\Theta)$ differs from \underline{E} , compute $\underline{Q}(A) = \underline{P}(P(A|\Theta))$. Let Φ denote the standard Normal distribution function.

Minimize $P(A|\theta)$ over $N(\theta + v, 1)$ distributions, giving $v = 1$ and $\underline{P}(A|\theta) = \Phi(\theta - 1)$. Then $\underline{P}(\underline{P}(A|\Theta))$ is minimized over $N(\mu, 1)$ priors by $\mu = 9$, giving $\underline{Q}(A) = 1 - \Phi(\sqrt{2}) = 0.079$. By symmetry $\bar{Q}(A) = \Phi(\sqrt{2}) = 0.921$. So the natural extension \underline{Q} of \underline{P} and $\underline{P}(\cdot|\Theta)$ is very imprecise, whereas the natural extension \underline{E} of \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ assigns precise probability to A .

8.4 Extension to posterior previsions

In the rest of this chapter we suppose that a sampling model and prior prevision are specified, and consider what they imply about posterior previsions. Their implications can again be described through their natural extension $\underline{E}(\cdot|\mathcal{X})$. Here coherent extension is not always possible. (That is already clear from the results of section 6.8.)

Although its importance has been exaggerated by Bayesians,¹ this is arguably the most important type of statistical extension, because it defines an updating strategy that is often useful for evaluating statistical evidence. The updating strategy involves assessing the sampling model $\underline{P}(\cdot|\Theta)$ and prior \underline{P} , constructing their natural extension $\underline{E}(\cdot|\mathcal{X})$, observing the outcome x of the statistical experiment, and adopting the lower prevision $\underline{E}(\cdot|x)$ as a model for updated beliefs about Θ . That is one way of drawing conclusions about Θ from the observation x , which is what is usually meant by statistical inference.

In statistical problems the sampling model is often firmly grounded in previous data or theory (as discussed in section 7.2), and assessment of a prior prevision \underline{P} presents the greater difficulties. But assessment of \underline{P} , which is based only on prior evidence, may be easier than direct assessment of the posterior prevision, based on all the statistical and prior evidence together. In such cases the updating strategy is useful because it separates the statistical evidence from the prior evidence.²

We assume that the sampling model $\underline{P}(\cdot|\Theta)$ is defined on a space \mathcal{F} of measurable gambles satisfying assumptions 7.3.2. There are several cases to consider, depending on the domain of the prior prevision \underline{P} . In the first case, \underline{P} is defined also on \mathcal{F} and represents prior beliefs about θ and x jointly. Then the natural extension $\underline{E}(\cdot|\mathcal{X})$ is determined by \underline{P} alone through the generalized Bayes rule, and \mathcal{X} -conglomerability of \underline{P} is necessary and sufficient for the extension to be coherent, just as in section 6.8. In the second case, \underline{P} represents prior beliefs about θ alone and is defined on a space \mathcal{K} satisfying assumptions 7.6.1. If the sampling model $\underline{P}(\cdot|\Theta)$ is linear then \underline{P} has a unique extension to \mathcal{F} that is coherent with $\underline{P}(\cdot|\Theta)$, and this case reduces to the first one. These two cases are treated in this section. The

remaining case, where \underline{P} is defined on \mathcal{K} and $\underline{P}(\cdot|\Theta)$ is non-linear, is treated in section 8.5.

Suppose first that \underline{P} and $\underline{P}(\cdot|\Theta)$ are both specified on \mathcal{F} and are coherent. Theorem 6.8.2 says that the natural extension of \underline{P} alone to previsions conditional on \mathcal{X} is defined by the generalized Bayes rule. The following lemma says that this also defines the natural extension of \underline{P} and $\underline{P}(\cdot|\Theta)$.

8.4.1 Lemma

Suppose that \underline{P} and $\underline{P}(\cdot|\Theta)$ are both specified on \mathcal{F} , satisfy assumptions 7.3.2, and are coherent. Then their natural extension $\underline{E}(\cdot|\mathcal{X})$ is given, for Y in \mathcal{F} , by the generalized Bayes rule $\underline{E}(Y|x) = \max\{\beta: \underline{P}(\{x\}(Y - \beta)) \geq 0\}$ when $\underline{P}(\{x\}) > 0$, and by $\underline{E}(Y|x) = \inf\{Y(\theta, x): \theta \in \Theta\}$ when $\underline{P}(\{x\}) = 0$.³

The natural extension $\underline{E}(\cdot|\mathcal{X})$ therefore depends only on \underline{P} , and not on the sampling model $\underline{P}(\cdot|\Theta)$. By the results of section 6.8, $\underline{E}(\cdot|\mathcal{X})$ is coherent with \underline{P} if and only if \underline{P} is \mathcal{X} -conglomerable, and otherwise there is no posterior prevision that is coherent with \underline{P} . It is not immediately obvious that \mathcal{X} -conglomerability suffices for $\underline{E}(\cdot|\mathcal{X})$ to be coherent with both \underline{P} and $\underline{P}(\cdot|\Theta)$, but that is confirmed by the following.

8.4.2 Posterior extension theorem

Suppose that \underline{P} and $\underline{P}(\cdot|\Theta)$ are defined on \mathcal{F} , satisfy assumptions 7.3.2, and are coherent. There is a posterior prevision $\underline{P}(\cdot|\mathcal{X})$ on \mathcal{F} such that \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent if and only if \underline{P} is \mathcal{X} -conglomerable. In that case the minimal coherent posterior is the natural extension $\underline{E}(\cdot|\mathcal{X})$ of \underline{P} , and any coherent posterior must agree with $\underline{E}(\cdot|x)$ whenever $\underline{P}(\{x\}) > 0$.

Proof. If \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent then \underline{P} and $\underline{P}(\cdot|\mathcal{X})$ are coherent, so \underline{P} is \mathcal{X} -conglomerable by Theorem 6.8.2. When $\underline{P}(\{x\}) > 0$, \underline{P} uniquely determines a coherent posterior through the generalized Bayes rule. It remains to prove that \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{E}(\cdot|\mathcal{X})$ are coherent provided \underline{P} is \mathcal{X} -conglomerable. (Minimality of $\underline{E}(\cdot|\mathcal{X})$ then follows by Theorems 6.8.2 or 8.1.2(e).) First verify, using 8.1.2(b), that $\underline{E}(\cdot|\mathcal{X})$ satisfies assumptions 7.3.2. Applying Theorem 7.9.1, \underline{P} , $\underline{P}(\cdot|\Theta)$ and $\underline{E}(\cdot|\mathcal{X})$ are coherent if and only if they are pairwise coherent. Now \underline{P} and $\underline{P}(\cdot|\Theta)$ are coherent by assumption. Assuming that \underline{P} is \mathcal{X} -conglomerable, \underline{P} and $\underline{E}(\cdot|\mathcal{X})$ are coherent by 6.8.2. By Theorem 7.3.6, we need to show that $\underline{P}(\cdot|\Theta)$ and $\underline{E}(\cdot|\mathcal{X})$ satisfy axiom S5. We use the following.

Lemma. If \underline{P} and $\underline{E}(\cdot|\mathcal{X})$ are coherent, $\bar{E}(Y|x) < 0$ for all $x \in A$, and $\underline{P}(\{y\}) > 0$ for some $y \in A$, then $\bar{P}(AY) < 0$.⁴

To verify S5, suppose that S5(i) fails. Then there are $Y \in \mathcal{F}$ and $\theta_0 \in \Theta$

such that $\bar{P}(Y|\theta_0) > 0$, $\underline{P}(Y|\theta) > 0$ for all other θ in $S(Y)$, $\bar{E}(Y|x) < 0$ for all $x \in T(Y)$. If $\underline{P}(\{x\}) = 0$ for all $x \in T(Y)$ then $\bar{E}(Y|x) = \sup Y(\cdot, x) < 0$ so that $Y \leq 0$, contradicting $\bar{P}(Y|\theta_0) > 0$. Suppose that $\underline{P}(\{y\}) > 0$ for some $y \in T(Y)$. Apply the lemma with $A = T(Y)$ to show that $\bar{P}(Y) = \bar{P}(AY) < 0$. But $\bar{P}(Y) \geq \bar{P}(\{\theta_0\}Y) + \underline{P}(\{\theta_0\}^cY) \geq \bar{P}(\{\theta_0\})\bar{P}(Y|\theta_0) + \inf \underline{P}(\{\theta_0\}^cY|\Theta) \geq 0$, a contradiction. S5(ii) can be verified in a similar way, using the lemma with $A = T(Y) \cap \{x_0\}^c$. ◆

This theorem closely resembles Theorem 6.8.2. The natural extension $\underline{E}(\cdot|\mathcal{X})$ of \underline{P} and $\underline{P}(\cdot|\Theta)$ is the same as the natural extension of \underline{P} alone, and it is coherent with \underline{P} and $\underline{P}(\cdot|\Theta)$ whenever it is coherent with \underline{P} alone. Thus $\underline{P}(\cdot|\Theta)$ provides no extra information to constrain posterior beliefs.⁵

Next consider the case where \underline{P} is defined on the space \mathcal{K} of \mathcal{C} -measurable gambles, representing prior beliefs about θ alone, and also $P(\cdot|\Theta)$ is linear. Let \underline{E} denote the unconditional natural extension of \underline{P} and $P(\cdot|\Theta)$ to \mathcal{F} , defined by $\underline{E}(Y) = \underline{P}(P(Y|\Theta))$. By section 6.7.3, \underline{E} is the unique extension of \underline{P} to \mathcal{F} that is coherent with $P(\cdot|\Theta)$. The natural extension $\underline{E}(\cdot|\mathcal{X})$ of \underline{P} and $P(\cdot|\Theta)$ is simply the natural extension of \underline{E} .⁶ Hence this case reduces to the previous one, with \underline{E} replacing \underline{P} .

8.4.3 Corollary

Suppose that \underline{P} is defined on \mathcal{K} , $P(\cdot|\Theta)$ is defined on \mathcal{F} and linear, and they satisfy assumptions 7.3.2 and 7.6.1. Let \underline{E} be their unique coherent extension to an unconditional revision on \mathcal{F} . There is a posterior revision on \mathcal{F} that is coherent with \underline{P} and $P(\cdot|\Theta)$ if and only if \underline{E} is \mathcal{X} -conglomerable. In that case the minimal coherent posterior is the natural extension $\underline{E}(\cdot|\mathcal{X})$ of \underline{E} , defined in Lemma 8.4.1.

Proof. \underline{P} and $P(\cdot|\Theta)$ are automatically coherent. If \underline{E} is \mathcal{X} -conglomerable then \underline{E} , $P(\cdot|\Theta)$ and $\underline{E}(\cdot|\mathcal{X})$ are coherent by the previous theorem, so \underline{P} , $P(\cdot|\Theta)$ and $\underline{E}(\cdot|\mathcal{X})$ are coherent. Conversely, suppose \underline{P} , $P(\cdot|\Theta)$, $\underline{P}(\cdot|\mathcal{X})$ are coherent. By the unconditional extension theorem 8.1.8, \underline{P} can be coherently extended to \mathcal{F} , and the extension must agree with \underline{E} to be coherent with $P(\cdot|\Theta)$. Thus \underline{E} , $P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent. By the previous theorem, \underline{E} must be \mathcal{X} -conglomerable and $\underline{P}(\cdot|\mathcal{X})$ must dominate $\underline{E}(\cdot|\mathcal{X})$. ◆

In both the cases examined, the necessary and sufficient condition for the existence of a coherent posterior is \mathcal{X} -conglomerability, either of \underline{P} (when specified on \mathcal{F}) or of its unique coherent extension \underline{E} . We argued in section 6.8 that conglomerability is a requirement of rationality. On the other hand, \mathcal{X} -conglomerability is sufficient to guarantee the existence of coherent posteriors. Recall that any of the following conditions is sufficient

8.4 EXTENSION TO POSTERIOR PREVISIONS

for \mathcal{X} -conglomerability of \underline{P} or \underline{E} : (a) \mathcal{X} is finite; (b) $\underline{E}(\{x\}) = 0$, or $\underline{P}(\{x\}) = 0$, for all x in \mathcal{X} ;⁷ (c) \underline{E} or \underline{P} is a lower envelope of countably additive linear previsions;⁸ or (d) each $\underline{P}(\cdot|\theta)$ is a countably additive linear revision, and \underline{P} (on \mathcal{K}) is a lower envelope of countably additive linear previsions.

8.4.4 The likelihood function

In the rest of this section we concentrate on the case in Corollary 8.4.3, where \underline{P} is specified on \mathcal{K} , $P(\cdot|\Theta)$ is linear on \mathcal{F} , and \underline{E} is their natural extension to \mathcal{F} . We first show that $\underline{E}(\cdot|x)$ depends on the sampling model only through the observed likelihood function.

The likelihood function generated by an observation x is the function L_x on Θ defined by $L_x(\theta) = \underline{P}(\{x\}|\theta)$. This function is bounded and \mathcal{C} -measurable by assumptions 7.3.2, so L_x can be regarded as a gamble in \mathcal{K} .

We have $\underline{E}(\{x\}) = \underline{P}(P(\{x\}|\Theta)) = \underline{P}(L_x)$. When $\underline{P}(L_x) = 0$, the natural extension $\underline{E}(\cdot|x)$ is vacuous. When $\underline{P}(L_x) > 0$, the natural extension is defined by the generalized Bayes rule. Now suppose that Y is in \mathcal{K} .⁹ Then $\underline{E}(\{x\}(Y - \beta)) = \underline{P}(P(\{x\}(Y - \beta)|\Theta)) = \underline{P}((Y - \beta)L_x)$. Thus the generalized Bayes rule may be written as $\underline{P}((Y - \underline{E}(Y|x))L_x) = 0$, and this uniquely determines $\underline{E}(Y|x)$ whenever $Y \in \mathcal{K}$ and $\underline{P}(L_x) > 0$.¹⁰

This shows that the natural extension $\underline{E}(\cdot|x)$ depends on a linear sampling model only through the likelihood function generated by x . (Indeed, it is clear that any positive multiple of L_x will give the same posterior.) Inferences about Θ that depend only on $\underline{E}(\cdot|x)$ will automatically obey the likelihood principle (section 8.6). Moreover, if $\underline{P}(L_x) > 0$ then the natural extension $\underline{E}(\cdot|x)$ is the unique coherent posterior, and any coherent inferences about Θ must obey the likelihood principle.

When L_x is a constant positive likelihood function, for example, the generalized Bayes rule gives $\underline{E}(Y|x) = \underline{P}(Y)$ for all $Y \in \mathcal{K}$. In that case the observation x is uninformative, in the sense that the posterior $\underline{E}(\cdot|x)$ agrees with the prior \underline{P} .

The fact that the posterior prevision depends on the sampling model only through the observed likelihood function can greatly simplify problems in which it is difficult to specify what other observations might have been made, or to assign them sampling probabilities. In such cases it suffices to assess the sampling probabilities $\underline{P}(\{x\}|\theta)$ of the actual observation x , after it is observed, without detailed consideration of other possibilities.

8.4.5 Linear posteriors

When is the posterior $\underline{E}(\cdot|x)$ a linear prevision on \mathcal{F} ? Sufficient conditions for linearity are that the prior \underline{P} and sampling models $P(\cdot|\Theta)$ are linear,

and x has positive prior probability $P(L_x)$. In that case the generalized Bayes rule reduces to Bayes' rule, $E(Y|x) = P(YL_x)/P(L_x)$ for $Y \in \mathcal{K}$.¹¹

It is obviously necessary for linearity of $\underline{E}(\cdot|x)$ that x have positive prior probability. We will see in section 8.5.4 that it is also necessary that the upper and lower likelihood functions $\bar{P}(\{\{x\}\}|\Theta)$ and $\underline{P}(\{\{x\}\}|\Theta)$ essentially agree, so that the likelihood function is essentially precise. Assuming that L_x is precise and bounded away from zero, Theorem 8.2.7 shows that a linear posterior $E(\cdot|x)$ determines a unique linear prior P by coherence. Thus, assuming the (upper) likelihood function is bounded away from zero, it is necessary and sufficient for linearity of the posterior $\underline{E}(\cdot|x)$ that the likelihood function is essentially precise and the prior is linear. Imprecise priors or imprecise likelihood functions generate imprecise posteriors.

8.4.6 Standard Bayesian models

Suppose P and $P(\cdot|\Theta)$ are linear previsions satisfying the conditions of the standard Bayesian model (Definition 7.7.1). The likelihood function is $L_x(\theta) = f(x|\theta)v(\{x\})$ and the observation x has prior probability $P(L_x) = q(x)v(\{x\})$. Provided both $q(x)$ and $v(\{x\})$ are positive, the natural extension agrees with the standard Bayesian posterior defined in 7.7.1(c), and the unique posterior prevision coherent with P and $P(\cdot|\Theta)$. For example, the beta posterior defined in Example 7.7.3 is the natural extension of the beta prior and binomial sampling model, and this is the unique posterior that is coherent with the prior and sampling model.

In contrast, consider the Normal model 7.8.3, where the sampling distributions are $N(\theta, 1)$ and the prior for θ is $N(0, 1)$. Each x has prior probability zero so the natural extension $\underline{E}(\cdot|\mathcal{X})$ is vacuous. Compare that with the standard Bayesian posteriors $P(\cdot|x)$, which are precise Normal $(\frac{1}{2}x, \frac{1}{2})$ distributions.¹² The standard Bayesian posteriors may be natural or justifiable in particular problems but they are not necessary for coherence. To justify them You must provide further information beyond that contained in the prior and sampling model.¹³

8.4.7 Vacuous posteriors

The natural extension $\underline{E}(\cdot|x)$ is vacuous whenever x has zero prior probability, and whenever the prior P is vacuous on \mathcal{K} (by Example 7.6.6).¹⁴ A vacuous prior, together with a sampling model, cannot generate any non-trivial beliefs or conclusions about Θ .

A stronger result can be obtained. Let A be any subset of Θ with prior probability $P(A) = 0$. Then $\underline{P}(AL_x) = 0$. When $\underline{P}(L_x) > 0$ the generalized

8.4 EXTENSION TO POSTERIOR PREVISIONS

Bayes rule gives $\underline{E}(A|x) = 0$,¹⁵ and this holds also when $\underline{P}(L_x) = 0$ since $\underline{E}(\cdot|x)$ is vacuous. Thus zero prior probability $\underline{P}(A)$ leads to zero posterior probability $\underline{E}(A|x)$ for every possible x . If You are initially unwilling to bet on A and You update by natural extension then You will remain unwilling to bet on A , whatever is observed. (Indeed, that holds for any coherent updated previsions, provided the observation has positive prior probability.)

If You can envisage an observation x that would lead You to bet on A at some odds, so that $\underline{P}(A|x) > 0$, then You should be willing to bet on A before making the observation, so that $\underline{P}(A) > 0$. Indeed You can use Your assessments of $\underline{P}(A|x)$ and $\underline{P}(L_x)$ to assess a positive prior probability $\underline{P}(A) \geq \underline{P}(L_x)\underline{P}(A|x)$. Also the theory or data on which the sampling model is based will usually provide some evidence about the parameter value. This may often be rather weak evidence, but sufficient to justify non-vacuous prior probabilities. So You should usually be able to assess positive prior probabilities $\underline{P}(A)$, at least for open sets A .¹⁶

8.4.8 Lower envelopes

Since the prior P is coherent, it is the lower envelope of its class of dominating linear previsions $\mathcal{M}(P)$. The generalized Bayes rule can be expressed in terms of $\mathcal{M}(P)$ as

$$\underline{E}(Y|x) = \min \{P(YL_x)/P(L_x) : P \in \mathcal{M}(P)\}, \text{ whenever } Y \in \mathcal{K} \text{ and } \underline{P}(L_x) > 0.\quad ^{17}$$

This formula is a simple generalization of Bayes' rule $E(Y|x) = P(YL_x)/P(L_x)$ to imprecise priors. It is often useful for computing the posterior $\underline{E}(\cdot|x)$. To do so it suffices to apply Bayes' rule to the extreme points of the prior class $\mathcal{M}(P)$, and take $\underline{E}(\cdot|x)$ to be the lower envelope of the resulting linear posteriors. To see that $\underline{E}(Y|x) = P(YL_x)/P(L_x)$ for some P that is an extreme point of $\mathcal{M}(P)$, apply Theorem 3.6.2(c) to the gamble $X = (Y - \underline{E}(Y|x))L_x$ to show that there is an extreme point P with $P(X) = \underline{P}(X)$. But $\underline{P}(X) = 0$ by the GBR (8.4.4). Hence $P(X) = 0$, which reduces to $\underline{E}(Y|x) = P(YL_x)/P(L_x)$ by linearity of P .

8.4.9 Beta-binomial example

Consider the imprecise beta-binomial model (Example 7.8.2), for which $P(\cdot|\theta)$ are binomial (n, θ) distributions and P is the lower envelope of a class of beta (s_γ, t_γ) distributions over an index set Γ . The likelihood function is $L_x(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$. When $\underline{P}(L_x) > 0$ there is a unique coherent posterior, and this must be the lower envelope of beta (s'_γ, t'_γ) posteriors

defined in Example 7.8.2. Now $\underline{P}(L_x) = \inf \{P_\gamma(L_x) : \gamma \in \Gamma\}$ where P_γ is a beta (s_γ, t_γ) distribution. We compute

$$\underline{P}_x(L_x) = \binom{n}{x} b(s'_y, t'_y)/b(s_y, t_y) \quad \text{where} \quad b(s, t) = \int_0^1 \theta^{st-1} (1-\theta)^{s(1-t)-1} d\theta.$$

It follows that $\underline{P}(L_x) > 0$ for each x provided s_y is bounded away from zero and t_y is bounded away from zero and one, over $\gamma \in \Gamma$. In that case the lower envelope of beta (s'_y, t'_y) posteriors is the unique coherent posterior.

However, the near-ignorance priors in section 5.3 have $0 < t < 1$, so $\underline{P}(L_x) = 0$ for all x in \mathcal{X} . Their natural extension $\underline{E}(\cdot|\mathcal{X})$ is therefore vacuous. The non-vacuous posteriors defined in section 5.3 are coherent, but not uniquely coherent.¹⁸

8.5 Posterior for imprecise sampling models

Next we examine the posteriors generated by natural extension of prior previsions and imprecise sampling models. The case in which the prior \underline{P} is specified on \mathcal{F} , describing beliefs about θ and x jointly, has already been covered by Theorem 8.4.2. We assume here that \underline{P} is specified on \mathcal{K} , describing prior beliefs about θ alone.

If \underline{P} has a unique extension to \mathcal{F} that is coherent with the sampling model, as when the sampling model is linear, then there is no loss of information in specifying \underline{P} on \mathcal{K} . If \underline{P} has many coherent extensions to \mathcal{F} , it is preferable to specify \underline{P} on \mathcal{F} rather than on \mathcal{K} , to provide more information about beliefs. The problem considered here therefore seems less important than the case of Theorem 8.4.2. However, it might be relatively easy to assess prior beliefs about θ alone, but difficult to assess beliefs about (θ, x) jointly except by natural extension. The problem treated here is therefore of some interest, especially as the posterior turns out to depend on the imprecise sampling model only through the upper and lower sampling probabilities of the observation. These can be regarded as imprecise (upper and lower) likelihood functions.

The following theorem characterizes the minimal coherent posterior in terms of coherent extensions of the prior to \mathcal{F} .

8.5.1 Posterior extension theorem

Suppose that \underline{P} is defined on \mathcal{K} , $\underline{P}(\cdot|\Theta)$ is defined on \mathcal{F} , and they satisfy assumptions 7.3.2 and 7.6.1. There is a posterior prevision on \mathcal{F} that is coherent with \underline{P} and $\underline{P}(\cdot|\Theta)$ if and only if \underline{P} has an extension \underline{Q} to \mathcal{F} that is coherent with $\underline{P}(\cdot|\Theta)$ and \mathcal{X} -conglomerable. In that case the minimal

coherent posterior is determined by the minimal extension \underline{Q} , through the generalized Bayes rule (8.4.1).¹

Proof. If there is such a \underline{Q} , by Theorem 8.4.2 there is a posterior that is coherent with \underline{Q} and $\underline{P}(\cdot|\Theta)$, and hence with \underline{P} and $\underline{P}(\cdot|\Theta)$. Conversely, if \underline{P} , $\underline{P}(\cdot|\Theta)$, $\underline{P}(\cdot|\mathcal{X})$ are coherent they can be extended to coherent \underline{Q} , $\underline{P}(\cdot|\Theta)$, $\underline{P}(\cdot|\mathcal{X})$ by the unconditional extension theorem, and \underline{Q} is \mathcal{X} -conglomerable since it is coherent with some posterior. ◆

Thus we are led to examine which extensions of \underline{P} to \mathcal{F} are coherent with $\underline{P}(\cdot|\Theta)$ and also \mathcal{X} -conglomerable. The most obvious candidate is the natural extension of \underline{P} and $\underline{P}(\cdot|\Theta)$, defined by $\underline{E}(Y) = \underline{P}(\underline{P}(Y|\Theta))$ for all $Y \in \mathcal{F}$. Then \underline{E} is coherent with $\underline{P}(\cdot|\Theta)$ by Theorem 6.7.2, so the only problem is whether \underline{E} is \mathcal{X} -conglomerable.

8.5.2 Corollary

Under the assumptions of Theorem 8.5.1, let \underline{E} be the natural extension of \underline{P} and $\underline{P}(\cdot|\Theta)$ to \mathcal{F} , and let $\underline{E}(\cdot|\mathcal{X})$ be the natural extension of \underline{E} defined by 8.4.1.² Provided \underline{E} is \mathcal{X} -conglomerable, $\underline{E}(\cdot|\mathcal{X})$ is the minimal coherent posterior of \underline{P} and $\underline{P}(\cdot|\Theta)$.³

Recall that \underline{E} is \mathcal{X} -conglomerable when \mathcal{X} is finite, or when $\underline{E}(\{x\}) = 0$ for all x , or when \underline{P} and each $\underline{P}(\cdot|\theta)$ are lower envelopes of countably additive linear previsions.

The rest of this section is concerned with the natural extension $\underline{E}(\cdot|\mathcal{X})$. Note, however, that even if this is a coherent posterior for \underline{P} and $\underline{P}(\cdot|\Theta)$, there will usually be other coherent posteriors because there are several coherent extensions of \underline{P} to \mathcal{F} .⁴

8.5.3 Upper and lower likelihood functions

Define the **lower likelihood function** L_x and the **upper likelihood function** U_x generated by an observation x by $L_x(\theta) = \underline{P}(\{x\}|\theta)$ and $U_x(\theta) = \bar{\underline{P}}(\{x\}|\theta)$. Then L_x and U_x can be regarded as gambles in \mathcal{K} . They are imprecise or ‘robust’ likelihood functions. When the sampling model $P(\cdot|\Theta)$ is linear, $L_x = U_x$ is a precise likelihood function. Generally U_x and L_x are upper and lower envelopes of the precise likelihood functions $P(\{x\}|\Theta)$, where $P(\cdot|\Theta)$ dominates $\underline{P}(\cdot|\Theta)$.

The natural extension $\underline{E}(\cdot|x)$ depends on the sampling model only through the upper and lower likelihood functions generated by x . To see that, note that $\underline{E}(\{x\}) = \underline{P}(\underline{P}(\{x\}|\Theta)) = \underline{P}(L_x)$. Hence $\underline{E}(\cdot|x)$ is vacuous when $\underline{P}(L_x) = 0$. When $\underline{P}(L_x) > 0$ and $Y \in \mathcal{K}$, $\underline{E}(Y|x)$ is the unique value of β such that $0 = \underline{E}(\{x\}(Y - \beta)) = \underline{P}(\underline{P}(\{x\}(Y - \beta)|\Theta))$. Define the medial likelihood

function $Z(Y, \beta)$ by $Z(Y, \beta)(\theta) = L_x(\theta)$ if $Y(\theta) \geq \beta$, $Z(Y, \beta)(\theta) = U_x(\theta)$ if $Y(\theta) < \beta$. Then $\underline{P}(\{x\}(Y - \beta)|\theta) = (Y(\theta) - \beta)Z(Y, \beta)(\theta)$. Hence $\underline{E}(Y|x)$ is the unique value of β satisfying $\underline{P}((Y - \beta)Z(Y, \beta)) = 0$, another version of the **generalized Bayes rule**. Clearly $Z(Y, \beta)$ and $\underline{E}(\cdot|x)$ depend on $\underline{P}(\cdot|\Theta)$ only through the upper and lower likelihood functions U_x and L_x .

For example, if A is a subset of Θ and $\underline{P}(L_x) > 0$ then $Z(A, \beta) = AL_x + A^c U_x$ for $0 < \beta \leq 1$, so that $\underline{E}(A|x)$ is the unique value of β in the interval $[0, 1]$ such that $\underline{P}(AL_x - \beta(AL_x + A^c U_x)) = 0$. Useful bounds on $\underline{E}(A|x)$ are⁵

$$\begin{aligned}\underline{P}(AL)/\bar{P}(U) &\leq \underline{P}(AL)/(\underline{P}(AL) + \bar{P}(A^c U)) \leq \underline{E}(A|x) \\ &\leq \min\{\underline{P}(AL)/\underline{P}(L), \bar{P}(AU)/\bar{P}(U)\}.\end{aligned}$$

More generally, properties 6.3.5 give the following bounds for $\underline{E}(Y|x)$ when $Y \in \mathcal{K}$ and $\underline{P}(L_x) > 0$:

- (a) if $\underline{P}(YL) = 0$ then $\underline{E}(Y|x) = 0$
- (b) if $\underline{P}(YL) > 0$ then

$$\underline{P}(YL)/\bar{P}(U) \leq \underline{E}(Y|x) \leq \min\{\underline{P}(YL)/\underline{P}(L), \bar{P}(YU)/\bar{P}(U)\}$$

- (c) if $\underline{P}(YL) < 0$ then

$$\underline{P}(YL)/\underline{P}(L) \leq \underline{E}(Y|x) \leq \min\{\underline{P}(YL)/\bar{P}(U), \bar{P}(YU)/\underline{P}(L)\}.$$

There are corresponding inequalities for $\bar{E}(A|x)$ and $\bar{E}(Y|x)$.

8.5.4 Lower envelopes

By Theorem 6.7.4, \underline{E} is the lower envelope of linear previsions of the form $P(P(\cdot|\Theta))$, where P dominates \underline{P} and each $P(\cdot|\theta)$ dominates $\underline{P}(\cdot|\theta)$. For each X in \mathcal{F} , $\underline{E}(X)$ is achieved by some such $P(P(X|\Theta))$. Applying this to $X = \{x\}(Y - \underline{E}(Y|x))$, where $Y \in \mathcal{K}$, we see that $\underline{E}(Y|x) = \min\{P(YZ)/P(Z): P \in \mathcal{M}(\underline{P}), L_x \leq Z \leq U_x\}$, provided x has positive prior lower probability $\underline{P}(L_x) = \min\{P(L_x): P \in \mathcal{M}(\underline{P})\}$.⁶ In that case $\underline{E}(\cdot|x)$ is the lower envelope of all Bayesian posteriors generated by linear priors P in $\mathcal{M}(\underline{P})$ and precise likelihood functions Z lying between the upper and lower likelihood functions.⁷

By the results in section 8.5.3, the minimizing likelihood function Z has the form $Z(Y, \beta)$ for some β , and this takes only the extreme values $L_x(\theta)$ and $U_x(\theta)$. Also, by the argument in section 8.4.8, P can be taken to be an extreme point of $\mathcal{M}(\underline{P})$. So the natural extension $\underline{E}(\cdot|x)$ can be constructed as a lower envelope of Bayesian posteriors generated by precise likelihood functions of the simple form $Z(Y, \beta)$ and linear priors that are extreme points of $\mathcal{M}(\underline{P})$. When A is an event, for example, we obtain $Z = AL + A^c U$ and $\underline{E}(A|x) = \min\{P(AL)/(P(AL) + P(A^c U)): P \in \text{ext } \mathcal{M}(\underline{P})\}$.

8.5 POSTERIORS FOR IMPRECISE SAMPLING MODELS

The precision of the posteriors $\underline{E}(\cdot|x)$ clearly increases with the precision of the prior \underline{P} and of the likelihood intervals $[L(\theta), U(\theta)]$. We saw in section 8.4.5 that, assuming the likelihood function $L = U$ is precise, precision (linearity) of the posterior is essentially equivalent to precision of the prior. We now show that precision of the posterior probability $E(A|x)$ for just a single non-trivial event A implies that the likelihood function is essentially precise, in the sense that $\bar{P}(U - L) = 0$. For suppose that $\bar{P}(AU) > 0$, $\bar{P}(A^c U) > 0$ and $\underline{E}(A|x) = \bar{E}(A|x)$. The above formula for $\underline{E}(A|x)$ and corresponding formula for $\bar{E}(A|x)$ then give $0 < \underline{E}(A|x) < 1$ and $P(AL) = P(AU)$, $P(A^c L) = P(A^c U)$ for all P in $\mathcal{M}(\underline{P})$. Hence $P(L) = P(U)$ for all P in $\mathcal{M}(\underline{P})$, and $\bar{P}(U - L) = 0$. Thus imprecise likelihood functions generate imprecise posterior probabilities for all non-trivial events.⁸

8.5.5 Upper and lower odds

As a simple example of these results, suppose that $\Theta = \{\theta, \psi\}$ contains only two parameter values. Then the prior prevision \underline{P} is determined by the prior upper and lower probabilities of $\{\theta\}$, or by the prior **upper** and **lower odds** on Θ , which are defined to be $\bar{\rho} = \bar{P}(\{\theta\})/\underline{P}(\{\psi\})$ and $\rho = \underline{P}(\{\theta\})/\bar{P}(\{\psi\})$ respectively. Similarly, the posterior prevision $\underline{E}(\cdot|x)$ is determined by the posterior upper and lower odds $\bar{\rho}(x)$ and $\rho(x)$.

Also define the **upper** and **lower likelihood ratios** generated by x to be $\bar{\lambda}(x) = U_x(\theta)/L_x(\psi)$ and $\underline{\lambda}(x) = L_x(\theta)/U_x(\psi)$. (These agree with the usual likelihood ratio when the likelihood function is precise.) Then the posterior odds on θ are given by the simple formulas $\bar{\rho}(x) = \bar{\rho}\bar{\lambda}(x)$ and $\rho(x) = \rho\underline{\lambda}(x)$.

As a numerical example, suppose that $\mathcal{X} = \{x, y\}$ represents the two possible ways in which a thumbtack can land. Parameter ψ represents the hypothesis that each outcome has precise chance $\frac{1}{2}$. Parameter θ represents an hypothesis that the chances are only approximately $\frac{1}{2}$, but may vary between trials according to how the tack is thrown, say $\underline{P}(x|\theta) = 0.4$, $\bar{P}(x|\theta) = 0.6$. This sampling model gives upper and lower likelihood functions $L(\psi) = U(\psi) = 0.5$, $L(\theta) = 0.4$, $U(\theta) = 0.6$, generated by both x and y . The upper and lower likelihood ratios are therefore $\bar{\lambda}(x) = \bar{\lambda}(y) = 1.2$, $\underline{\lambda}(x) = \underline{\lambda}(y) = 0.8$.

Suppose also that Your prior beliefs about Θ are described by $\underline{P}(\{\theta\}) = 0.4$, $\bar{P}(\{\theta\}) = 0.5$, giving prior upper and lower odds $\bar{\rho} = 1$, $\rho = \frac{2}{3}$. We obtain posterior upper and lower odds $\bar{\rho}(x) = \bar{\rho}(y) = 1.2$, $\rho(x) = \rho(y) = 0.53$. From these we can compute the posterior upper and lower probabilities $\bar{E}(\{\theta\}|x) = \bar{\rho}(x)/(1 + \bar{\rho}(x)) = \frac{6}{11} \simeq 0.55$, $\underline{E}(\{\theta\}|x) = \rho(x)/(1 + \rho(x)) = \frac{8}{23} \simeq 0.35$. The posterior probabilities conditional on y are identical. The effect of the imprecise sampling model $\underline{P}(\cdot|\theta)$ is to produce posterior probabilities that are less precise than the prior probabilities.

The natural extension $E(\cdot|\mathcal{X})$ is coherent with \underline{P} and $\underline{P}(\cdot|\Theta)$ here, since \mathcal{X} is finite, but there are other coherent posteriors. To see that, write \underline{P} and $\underline{P}(\cdot|\Theta)$ as the lower envelopes of linear previsions defined by $P_1(\theta) = 0.5$, $P_1(x|\theta) = 0.4$, $P_1(x|\psi) = 0.5$, and $P_2(\theta) = 0.4$, $P_2(x|\theta) = 0.6$, $P_2(x|\psi) = 0.5$. These generate linear posteriors by Bayes' rule, $P_1(\theta|x) = \frac{4}{9} = P_2(\theta|x)$, $P_1(\theta|y) = \frac{6}{11}$, $P_2(\theta|y) = \frac{8}{23}$. By Theorem 7.1.6, the lower envelope $\underline{P}(\cdot|\mathcal{X})$ of the two linear posteriors is coherent with \underline{P} and $\underline{P}(\cdot|\Theta)$. Since $P(\cdot|x)$ is a precise prevision, this differs from the natural extension.⁹

8.6 The likelihood principle

In this section we consider the extent to which coherence supports the likelihood principle.¹ We noted in section 8.4.4 that the posterior prevision generated by natural extension of a prior and linear sampling model depends on the sampling model only through the observed likelihood function. That supports a version of the likelihood principle for discrete sample spaces. We will argue that the principle is less reliable when it is applied to continuous sample spaces.

Suppose that a statistical observation x is generated by one of the sampling models $P(\cdot|\theta)$. Knowledge of $P(\cdot|\Theta)$ and x provides some information about the unknown parameter θ . According to the likelihood principle, the only part of this information that is relevant to inferences about θ is the observed likelihood function L_x , defined by $L_x(\theta) = P(\{x\}|\theta)$. Here is a more careful statement of the principle.

8.6.1 Discrete likelihood principle

Suppose that two statistical experiments, modelled by linear sampling models $P_1(\cdot|\Theta)$ and $P_2(\cdot|\Theta)$, are envisaged to provide information about an unknown state θ . Here θ represents the same physical state in both models, the two sampling models are completely known except for θ , and the true state is known to belong to Θ . Let x_1 and x_2 be possible outcomes of the two experiments. Suppose that they generate likelihood functions $L_1(\theta) = P_1(\{x_1\}|\theta)$ and $L_2(\theta) = P_2(\{x_2\}|\theta)$ that are proportional, meaning that $L_1 = \lambda L_2$ for some positive λ , and not identically zero. Then the observations x_1 and x_2 should lead to the same beliefs and conclusions about Θ .² More specifically, $P_1(\cdot|x_1)$ and $P_2(\cdot|x_2)$, the posterior previsions after observing x_1 or x_2 , should be identical.

To illustrate the force of this principle, let θ be the chance that a thumbtack will land pin-up when it is tossed. Various experiments could be carried out to give information about θ . One experiment is to toss the thumbtack a predetermined number of times (n), and observe the number of pin-ups (m).

A second experiment is to toss the thumbtack until a predetermined number of pin-ups (m) occur, and observe the total number of tosses (n). If the two experiments result in the same outcomes (m, n), then the two likelihood functions are proportional. According to the discrete likelihood principle, all the information about θ from the experiment is contained in the observed likelihood function $L(\theta) \propto \theta^m(1 - \theta)^{n-m}$. Inferences about θ should depend only on the data (m, n), and not on how the data were generated.³

8.6.2 Justification of the discrete likelihood principle

The discrete likelihood principle is a consequence of coherence, provided the following assumptions are made:

1. Conclusions about Θ depend on the statistical evidence (i.e., the sampling model and observation) only through posterior beliefs about Θ after observing x .
2. Posterior beliefs about Θ can be modelled by a posterior prevision $\underline{P}(\cdot|x)$.⁴
3. $\underline{P}(\cdot|\mathcal{X})$ is coherent with $\underline{P}(\cdot|\Theta)$, and therefore coherent with some prior prevision \underline{P} .⁵
4. The appropriate \underline{P} does not depend on which experiment is performed.⁶
5. The prior lower probability of either observation is non-zero, i.e. $\underline{P}(L_1) > 0$.

Assuming conditions (3) and (5), the posterior prevision $P_1(\cdot|x_1)$ is uniquely determined by the prior prevision \underline{P} and the likelihood function L_1 , through the generalized Bayes rule (section 8.4.4). Using (4), any likelihood function L_2 that is proportional to L_1 yields the same posterior prevision $P_2(\cdot|x_2) = P_1(\cdot|x_1)$. So x_1 and x_2 lead to the same posterior prevision. Using (1) and (2), they must lead to the same beliefs and conclusions about Θ .

It does seem reasonable to require that statistical inferences should satisfy the first four conditions. (For justification, see especially section 2.11.1.) Condition (5) is redundant when the likelihood function L_1 is bounded away from zero, since then $\underline{P}(L_1) > 0$ for any coherent prior. That usually holds when both Θ and \mathcal{X} are finite. More generally, it is arguable that non-zero probabilities should be assigned to all possible observations in discrete sample spaces,⁷ and especially to those that actually occur. So the likelihood principle appears to be justified for discrete sample spaces.⁸

Now consider continuous sample spaces. The likelihood function, as previously defined, will typically be identically zero. However, it is usual to extend the likelihood principle to continuous spaces by defining the likelihood function in terms of density functions. As in Definition 7.7.1, suppose that each $P(\cdot|\theta)$ is a countably additive, linear prevision on \mathcal{F} and

is absolutely continuous with respect to a σ -finite measure v . By the Radon–Nikodym theorem, each $P(\cdot|\theta)$ has a density function $f(\cdot|\theta)$ with respect to v . Define the likelihood function generated by an observation x to be $L_x(\theta) = f(x|\theta)$. The continuous version of the likelihood principle then asserts that observations which generate proportional likelihood functions provide the same information about Θ .

8.6.3 Continuous likelihood principle

Suppose that the hypotheses of 8.6.1 hold, except that the likelihood functions are defined in terms of density functions. Then possible observations x_1 and x_2 , which generate proportional likelihood functions concerning the same state θ , should lead to the same posterior prevision $P_1(\cdot|x_1) = P_2(\cdot|x_2)$, and to the same beliefs and conclusions about Θ .

The immediate difficulty with this principle is that the density functions $f(\cdot|\theta)$, defined as Radon–Nikodym derivatives, are not uniquely determined by $P(\cdot|\theta)$ and v . Any version of $f(\cdot|\theta)$ can be modified on a set of v -measure zero. The lack of uniqueness is far from negligible when the sample space is continuous. Then typically $v(\{x\}) = 0$, so $f(x|\theta)$ is wholly undetermined, and the likelihood function generated by x is wholly undetermined. The measure-theoretic approach using Radon–Nikodym derivatives is therefore inadequate to determine a useful likelihood function.⁹

8.6.4 Continuous sample spaces as idealizations of discrete spaces

A different approach is needed to justify a unique version of the density functions and hence a unique likelihood function. The most natural approach is to regard the continuous sample space as a mathematical idealization of a discrete experiment. That is natural because practical measurements, at least those that are reported numerically, are rounded or grouped. The measurement actually has limited precision but it is regarded as drawn from a continuum of ‘possible’ measurements. The idealization can be useful for mathematical tractability (continuous models are often easier to manipulate), or because the degree of imprecision or rounding is unknown. We argued in section 6.10 that, in order to draw sensible conclusions from continuous data, we need to know what kind of measurement imprecision is involved.¹⁰ That is, we need to know what kind of discrete observations were actually made. (The degree of imprecision may not be crucial, but the way in which the imprecision depends on the unknown state θ is crucial.)

Suppose, for example, that the sample space \mathcal{X} is the real line, and the reported observation x represents the event that an underlying continuous variable belongs to the interval $B(x, \delta) = (x - \delta, x + \delta)$.¹¹ Possibly the

imprecision δ is unknown, but we assume that it does not depend on the parameter θ . If we knew δ , we would adopt $L(\theta) = P(B(x, \delta)|\theta)$ as the observed likelihood function, and that would lead to a unique coherent posterior prevision. Equivalently we could adopt the proportional likelihood function $P(B(x, \delta)|\theta)/2\delta$. This quantity will converge to a limit $f(x|\theta)$ as δ tends to zero, provided the cumulative distribution function for $P(\cdot|\theta)$ is differentiable at x . In that case the limit $f(x|\theta)$ defines the appropriate density function and continuous likelihood function. The justification is that, provided δ is small, the continuous likelihood function is a good approximation to the appropriate discrete likelihood function and will generate approximately the same posterior prevision.

The continuous likelihood principle therefore seems justified, via the discrete likelihood principle, in those cases where:

1. The reported value x represents a discrete observation $B(x, \delta)$ which occurs with positive sampling probabilities $P(B(x, \delta)|\theta)$.
2. The values $f(x|\theta)$ used to define the continuous likelihood function are limits (uniform in θ) of normalized probabilities $P(B(x, \delta)|\theta)/v(B(x, \delta))$ as $\delta \rightarrow 0$, where v is a suitable normalizing measure.¹²
3. The imprecision δ is sufficiently small and does not depend on θ .¹³

The density $f(x|\theta)$ defined above is simply the derivative at x of the cumulative distribution function for $P(\cdot|\theta)$. Call $f(\cdot|\theta)$ a **regular** density function for $P(\cdot|\theta)$ if $f(x|\theta)$ agrees with the derivative of the distribution function wherever this exists. If the distribution is absolutely continuous with respect to Lebesgue measure then the derivative exists except on a set of Lebesgue measure zero. For most standard sampling models the derivative exists at all but finitely many points.

In many statistical problems it is natural to use the regular density functions to define a unique likelihood function. However, Borel’s paradox (Example 6.10.1) and the next example show that the regular density functions can sometimes generate the wrong posterior distributions, even when they approximate the probabilities of the discrete observation as in (2). Assumption 3, that the imprecision does not depend on θ , is essential.

8.6.5 Inferences from uniform distributions¹⁴

Let both the sample space \mathcal{X} and parameter space Θ be the open interval $(0, 1)$. Suppose that the unknown parameter θ represents the number of units on some scale of measurement B equal to one unit on a different scale A . It is prior knowledge that $0 < \theta < 1$. A quantity z is measured on scale A and reported on scale B as $x = \theta z$. Assume that z has uniform distribution on $(0, 1)$. The observation x then has a uniform distribution on $(0, \theta)$, so that

the sampling model $P(\cdot|\Theta)$ is defined by $P(Y|\theta) = \theta^{-1} \int_0^\theta Y(\theta, x) dx$ for Borel-measurable gambles Y . The regular density function is $f(x|\theta) = \theta^{-1}$ for $0 < x < \theta$, $f(x|\theta) = 0$ for $\theta < x < 1$, with $f(x|\theta)$ undefined at $x = \theta$ as the distribution function is not differentiable there.

To simplify inferences from this model, suppose also that Your prior beliefs about Θ are represented by a uniform distribution on $(0, 1)$. (Nothing hinges on this choice.) Suppose now that You observe $x = 0.2$. What conclusions should You draw about θ ?

According to the approach developed in this chapter, we can draw no conclusions from the information given. Since the observation $x = 0.2$ has zero prior probability, the natural extension $E(\cdot|x=0.2)$ of the prior and sampling model is vacuous, and any posterior $P(\cdot|x=0.2)$ is coherent with them. There may be two different experiments that can both be described by this model and which yield the same outcome $x_1 = x_2 = 0.2$, but for which two different posteriors $P_1(\cdot|x_1=0.2)$ and $P_2(\cdot|x_2=0.2)$ can be justified by further information.

Contrary to this, the continuous likelihood principle asserts that such observations x_1 and x_2 should lead to the same posterior prevision. If we adopt the regular density functions, the continuous version of Bayes' rule determines the posterior as the linear prevision with density function $g(\theta|x=0.2) = (\theta \log 5)^{-1}$ on the interval $(0.2, 1)$ and zero density elsewhere. Call this the **regular Bayesian posterior**. Thus Bayes' rule determines a unique linear posterior, whereas coherence places no constraints on the posterior prevision.

To examine which of these inferences is appropriate, consider two ways in which the precise observations x_1, x_2 might be idealizations of discrete observations.

8.6.6 First model

Suppose first that the quantity $x = \theta z$ is reported by rounding to within $\theta\delta$. The reported quantity x_1 is the shortest terminating decimal that differs from x by less than $\theta\delta$.¹⁵ Assuming that the constant δ is less than 0.05, the observation $x_1 = 0.2$ represents an interval of values $C(\delta, \theta) = (0.2 - \theta\delta, 0.2 + \theta\delta)$ for the precise quantity x . The imprecision $\theta\delta$ concerning x depends on the unknown parameter θ , but the imprecision concerning $z = x/\theta$ is δ , independent of θ . This model is appropriate if there is constant imprecision in the original measurement of z on scale A , which generates imprecision proportional to θ on scale B through the scale transformation $x = \theta z$.

The appropriate discrete likelihood function is $L_1(\theta) = P(C(\delta, \theta)|\theta)$, which is 2δ if $0.2/(1-\delta) \leq \theta < 1$, zero if $0 < \theta \leq 0.2/(1+\delta)$, and intermediate between 0 and 2δ for other values of θ . Applying the discrete version of Bayes' rule, the density of the unique coherent posterior (for uniform prior)

is a positive constant for $0.2/(1-\delta) \leq \theta < 1$ and zero for $0 < \theta \leq 0.2/(1+\delta)$. As δ tends to zero, the posterior distribution approaches the uniform distribution on the interval $(0.2, 1)$. In effect, the observation x_1 tells us only that θ is larger than 0.2.¹⁶

Note that this model does converge to the uniform sampling model as δ tends to zero, and in fact its limiting densities are the regular uniform densities. Any x in \mathcal{X} belongs to some discrete event $B(x, \theta\delta)$ containing all precise values that are rounded to the same x_1 as x is. (The reported quantity x_1 does represent an event of this form.) We have $v(B(x, \theta\delta)) < 2\theta\delta$, where v denotes Lebesgue measure, and $P(B(x, \theta\delta)|\theta)/v(B(x, \theta\delta))$ converges to the regular density $f(x|\theta)$ as $\delta \rightarrow 0$, except at $x = \theta$. Thus the regular density functions can be 'justified' as limits of normalized discrete probabilities $P(B(x, \theta\delta)|\theta)$. Nevertheless, the regular Bayesian posteriors disagree with the uniform posteriors obtained as limits of discrete posteriors.¹⁷

8.6.7 Second model¹⁸

Suppose next that x is reported by rounding to within δ . The reported quantity x_2 is the shortest terminating decimal that differs from x by less than δ . Now the imprecision in measurements on scale B is constant, but the imprecision in the original measurement z depends strongly on θ . Again the uniform sampling model and regular density functions are obtained as limits of discrete probabilities. (The argument is exactly as for model 1, except that $\theta\delta$ is replaced by δ .)

The observation $x_2 = 0.2$ now represents the interval $D(\delta) = (0.2 - \delta, 0.2 + \delta)$ for the precise quantity x . The discrete likelihood function is $L_2(\theta) = P(D(\delta)|\theta)$, which is $2\delta/\theta$ if $0.2 + \delta \leq \theta < 1$, zero if $0 < \theta \leq 0.2 - \delta$, and intermediate for other values of θ . The discrete version of Bayes' rule gives a unique coherent posterior with density proportional to θ^{-1} for $0.2 + \delta \leq \theta < 1$ and zero density for $0 < \theta \leq 0.2 - \delta$. As δ tends to zero the posterior distribution approaches the regular Bayesian posterior.

The two models therefore lead to rather different posterior distributions. For example, the event that θ is less than 0.4 has posterior probability 0.25 under the first model and posterior probability 0.43 under the second model.

8.6.8 Conclusions

Each model might be reasonable in practice, depending on what is known about the precision of measurements on the two scales and the rounding of the reported observation. In the absence of such information, neither model can be ruled out and no precise posterior distribution is justified. In particular, the regular Bayesian posterior is unjustified without the extra information expressed in model 2. Of course other models, leading to

different posteriors, are also possible, and the vacuous posterior mandated by our theory (in the absence of further information) seems reasonable. A fixed sampling model, prior prevision and continuous observation x can lead to different conclusions depending on how x is an idealization of discrete observations. The continuous likelihood principle 8.6.3 therefore seems unreliable.

Advocates of the likelihood principle might object that the principle does not fail here, as it is recognized that the continuous sampling model is merely an idealization of the appropriate discrete model, to which the discrete likelihood principle applies. Careful statements of the likelihood principle emphasize that it is applicable only when the sampling model is assumed to be correct. However, it seems that continuous sampling models are never correct in the required way, since they are always idealizations of discrete experiments. The example illustrates that the likelihood principle cannot be applied reliably to continuous models, even to draw idealized or approximate conclusions from them, unless there is some information about how the continuous model approximates the actual observation.

Of course, this does not support the usual objections to the likelihood principle from frequentist statisticians, who argue that conclusions should depend on further information contained in the sampling model but not in the observed likelihood function. In the previous example the continuous sampling model provides no useful information. Statisticians of all schools seem to be agreed that a sampling model, prior distribution and observation together provide enough information to yield useful inferences.¹⁹ (Different schools would ignore different aspects of this information.) That seems doubtful, in view of the example, when the sample space is continuous.

The indeterminacy present in this example does occur in practical problems, especially where unknown transformations of scale are involved. The rounding of data may be influenced by the state θ , as in model 1. When data are reported there is often too little attention given to questions of rounding and precision. It might be suspected that continuous models are sometimes used in practice to avoid these questions, by treating the observations as completely precise. But that is unjustified, even if the degree of imprecision is extremely small. Continuous models lead to indeterminate conclusions unless some assumptions are made about the kind of imprecision involved.

8.6.9 Upper and lower likelihoods

Finally, consider an imprecise sampling model $\underline{P}(\cdot|\Theta)$. For any prior beliefs about Θ , the natural extension $\underline{E}(\cdot|x)$ of the prior and sampling model depends on the sampling model only through the upper and lower likelihood

8.6 THE LIKELIHOOD PRINCIPLE

functions $U_x(\theta) = \bar{P}(\{x\}|\theta)$ and $L_x(\theta) = \underline{P}(\{x\}|\theta)$. For a fixed prior, two observations x_1 and x_2 which generate proportional upper and lower likelihood functions (i.e., $U_1 = \lambda U_2$ and $L_1 = \lambda L_2$ for some positive λ) will determine the same posterior prevision by natural extension.²⁰ This suggests a generalization of the discrete likelihood principle: observations x_1 and x_2 (from discrete spaces) that generate proportional upper and lower likelihood functions should lead to the same posterior prevision and the same inferences about Θ .

The generalization is unreasonable, because a given prior prevision and imprecise sampling model may not determine a unique coherent posterior, even when the observation has positive prior probability. Consider the numerical example in section 8.5.5, with $\mathcal{X} = \{x, y\}$ and $\Theta = \{\theta, \psi\}$. Here x and y generate the same upper and lower likelihood functions, so the generalized likelihood principle would require the posteriors $\underline{P}(\cdot|x)$ and $\underline{P}(\cdot|y)$ to be the same. But this is not necessary for coherence, and it can be incompatible with coherence when You assess a joint prior prevision for (θ, x) . For example, define the prior \underline{P} on $\mathcal{L}(\Theta \times \mathcal{X})$ to be the lower envelope of P_1 and P_2 in Example 8.5.5. (So \underline{P} is coherent with $\underline{P}(\cdot|\Theta)$.) Then \underline{P} determines a unique coherent posterior $\underline{P}(\theta|x) = \bar{P}(\theta|x) = 0.44$, $\underline{P}(\theta|y) = 0.35$, $\bar{P}(\theta|y) = 0.55$. Here $P(\cdot|x)$ is a linear prevision but $\underline{P}(\cdot|y)$ is imprecise. You can draw sharper conclusions about Θ from x than from y , although x and y generate the same upper and lower likelihood functions.

CHAPTER 9

Structural judgements

This chapter is concerned with structural judgements, which express structural properties of beliefs such as independence or permutation-invariance. The aims are to:

1. define the most important types of structural judgement (independence, conditional independence, permutability and exchangeability) so that these concepts have simple behavioural interpretations;
2. consider the types of evidence that justify such judgements;
3. discuss the role of structural judgements in probability assessment, e.g. in constructing joint previsions from independent marginals;
4. compare our definitions of structural judgements with the definitions suggested by sensitivity analysis, to show that the difference in interpretation can lead to differences in practice.

Concepts of epistemic independence are defined for events and experiments in sections 9.1 and 9.2 respectively. Independence has the behavioural meaning that beliefs about one event or experiment would not change if You learned the outcome of another. Such judgements are based on evidence of unrelatedness or physical independence. When the probabilities of all outcomes are precise and positive, our definitions of independence are equivalent to the standard definitions involving factorization of joint probabilities. We give examples to show that, when some outcomes have probability zero, the standard definitions are too weak to capture the intuitive notion of independence. In the case of imprecise probabilities, our definitions are weaker than those suggested by sensitivity analysis. Events can be pairwise independent under a lower prevision but not under any of its dominating linear previsions.

In section 9.3 we consider what is implied about a joint experiment by assessments concerning two marginal experiments, together with a judgement that the marginals are independent. The results of Chapter 8 are used to construct a minimal product prevision by natural extension. This is compared with two different product previsions suggested by sensitivity analysis.

9.1 INDEPENDENT EVENTS

Judgements of permutability and exchangeability are discussed in sections 9.4 and 9.5. Permutability means that beliefs about a set of experiments are invariant under permutations of the experiments. Exchangeability means that You are willing to exchange any gamble for a permuted gamble. In the case of linear previsions, permutability and exchangeability are equivalent and agree with de Finetti's concept of exchangeability. For non-linear previsions, exchangeability is stronger than permutability. It is argued that a judgement of permutability is justified whenever the relevant evidence is symmetric in the experiments, even when there is little or no evidence, whereas exchangeability is a stronger judgement that requires substantial evidence about the physical structure of the experiments. For that and other reasons, de Finetti's attempt to eliminate aleatory probabilities from statistics, by 'reducing' them to epistemic probabilities through exchangeability, is unsuccessful.

Section 9.6 illustrates the role of structural judgements in constructing statistical models. We construct a robust version of the standard Bernoulli model for repetitions of a binary experiment, by introducing imprecision to allow for instability or indeterminacy in the process. The robust model leads to beliefs about future observations that are permutable but not exchangeable. There are limits to the precision with which we can estimate the Bernoulli parameter from a long series of observations.

Finally, section 9.7 outlines a general theory of structural judgements. The theory of elicitation in Chapter 4 can be extended to admit all types of structural judgements, by regarding them as structural constraints on a probability model. An argument against the sensitivity analysis interpretation is that it is not compatible, in general, with our behavioural interpretation of structural judgements.

9.1 Independent events

Judgements of independence have an important role in probability assessment. Often these are judgements that experiments are independent or conditionally independent, but first we examine the simpler case of independent events. We will say that two events are **epistemically independent** for You, or just **independent**, when Your beliefs about either event are unchanged by extra information that specifies whether or not the other event has occurred. That is, You regard the occurrence of either event as irrelevant (on Your current evidence) to beliefs about the other one. The formal definition is as follows.

9.1.1 Definition

Suppose that A and B are non-trivial events, and the following unconditional

and conditional probabilities are defined and coherent. Say that B is **irrelevant** to A when $\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A)$ and $\bar{P}(A|B) = \bar{P}(A|B^c) = \bar{P}(A)$. Say that A and B are (**epistemically independent**) when B is irrelevant to A and A is irrelevant to B .¹

Thus independence involves eight equalities amongst conditional and unconditional probabilities. Denoting the partitions $\{A, A^c\}$ and $\{B, B^c\}$ by \mathcal{A} and \mathcal{B} respectively, the eight equalities can be written most simply as $\underline{P}(C|\mathcal{B}) = \underline{P}(C)$ when $C \in \mathcal{A}$ and $\bar{P}(D|\mathcal{A}) = \bar{P}(D)$ when $D \in \mathcal{B}$.²

9.1.2 Grounds for independence judgements

Independence is defined as a property of the events and probability model P , $\underline{P}(\cdot|\mathcal{A})$, $\underline{P}(\cdot|\mathcal{B})$. However, judgements of independence are usually qualitative judgements, not based on a quantitative comparison of probabilities. Their main role is in constructing, rather than describing, probability models.³

According to the account of assessment strategies sketched in section 1.7, such judgements should be firmly grounded in evidence. What are the grounds for a judgement of independence? The most obvious ground is knowledge that the two events are unrelated, in the sense that there is no causal or evidential connection between them. The outcome of my next coin-toss is unrelated to a rise in next month's unemployment rate; one event cannot influence the other or provide evidence about the factors which influence it, and the outcome of either one should be regarded as irrelevant to the other. Such judgements are common and often intuitively obvious, but of course they rely on practical and theoretical knowledge concerning causal relationships.

It is not sufficient for epistemic independence that events be causally unrelated, in the sense that the occurrence of one has no causal influence on the other. Causal unrelatedness can be modelled through a concept of physical independence. Call two events **physically independent** when the occurrence or non-occurrence of one does not change the aleatory probability of the other. (That is, Definition 9.1.1 holds when the probabilities are aleatory rather than epistemic.) Unlike epistemic independence, physical independence is an objective property not depending on judgement or on evidence. Many of the models commonly used in statistics are based on hypotheses of physical independence.

Knowledge of causal unrelatedness and physical independence is sometimes, but not always, a ground for epistemic independence. If, for example, successive coin-tosses are physically independent but the bias of the coin is unknown, the outcome of one toss provides evidence about the bias and hence about the outcome of a second toss. The tosses are causally

9.1 INDEPENDENT EVENTS

unrelated but evidentially related, and they are physically independent but not epistemically independent. But if the bias of the coin is known then physical independence of the tosses does justify a judgement of epistemic independence. In that case the result of the first toss adds no relevant evidence concerning the second. Generally, by the principle of direct inference, knowledge of physical independence will justify a judgement of conditional epistemic independence, conditional on a partition that completely specifies the aleatory probabilities involved.

Knowledge of physical independence is not sufficient for epistemic independence, nor is it necessary. If You are completely ignorant about the chances of two events and about the physical dependence between them, then the events are epistemically independent (see 9.1.3).⁴

Qualitative judgements of independence are therefore natural in many contexts, based on an understanding of causal connections and relevance of evidence.⁵ They can be useful in several ways: (a) in decomposing an assessment problem into separate assessments, to be combined through independence (see section 9.3); (b) in simplifying evidence by eliminating any that is irrelevant to assessing particular probabilities; and (c) in imposing constraints (9.1.1) to make probability models more precise.

9.1.3 Logical independence

Suppose that You are completely ignorant about a possibility space Ω , so that all Your conditional and unconditional probabilities are vacuous. Then two non-trivial events A and B are epistemically independent for You if and only if they are **logically independent**, i.e. each of the four sets $A \cap B$, $A \cap B^c$, $A^c \cap B$, $A^c \cap B^c$ is non-empty. In such a case of complete ignorance, the occurrence of one event provides information only about those events it logically entails or with which it is logically inconsistent.

In general, epistemic independence implies logical independence provided the events A and B are not both 'essentially trivial'. (A is said to be essentially trivial when either $\bar{P}(A) = 0$ or $\bar{P}(A^c) = 0$).⁶

9.1.4 Standard definition of independence

Suppose that the specified probabilities are all precise. Independence of A and B then reduces to the four equations $\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A)$, $\bar{P}(B|A) = \bar{P}(B|A^c) = \bar{P}(B)$. Coherence implies that $\underline{P}(A \cap B) = \underline{P}(A)\underline{P}(B)$. Hence the **factorization property** $\underline{P}(A \cap B) = \underline{P}(A)\underline{P}(B)$ that constitutes the standard definition of independence is a consequence of our definition.

Provided the four events A , A^c , B , B^c each have positive probability, A

and B are independent if and only if $P(A \cap B) = P(A)P(B)$. (Since then the conditional probabilities are uniquely determined by the unconditional ones through Bayes' rule.) In that case our definition is equivalent to the standard definition and depends only on unconditional probabilities.

In general our definition of independence is stronger than the standard definition. For example, any event A with probability zero is independent of any other event B under the standard definition. If, say, B is disjoint from A and has positive probability, A and B cannot be independent under our definition because $P(B|A) = 0$ whereas $P(B) > 0$. (Let $B = A^c$, for instance; it is absurd to say that A and A^c are independent events.) As usual the standard approach disregards events of zero probability. That conflicts with our intuitive understanding of independence in the present case, and it leads to more serious difficulties in defining independent experiments (see section 9.2).

In the case of non-additive probabilities, the factorization property $P(A \cap B) = P(A)P(B)$ generalizes to $\underline{P}(BG(A)) = 0$, where as usual $G(A) = A - \underline{P}(A)$. The eight equations of the form $\underline{P}(CG(D)) = 0$, where C and D are chosen from the partitions $\{A, A^c\}$ and $\{B, B^c\}$, are necessary for independence of A and B , by the generalized Bayes rule. Provided A, A^c, B, B^c each have positive lower probability, these eight equations are also sufficient for independence. In that case, but not in general, independence can be defined in terms of unconditional previsions.

Using the equations $\underline{P}(BG(A)) = 0$, it can be shown that independent events A and B must satisfy the inequalities

$$\underline{P}(A)\underline{P}(B) \leq \underline{P}(A \cap B) \leq \underline{P}(A)\bar{P}(B) \leq \bar{P}(A \cap B) \leq \bar{P}(A)\bar{P}(B).^7$$

9.1.5 Sensitivity analysis definition of independence

When would a sensitivity analyst regard events as independent? Presumably he would regard independence as a property of, or a judgement about, an unknown linear revision P_t that represents Your true or ideal beliefs. He would then model Your beliefs by the class of all linear revisions which have the appropriate independence properties and satisfy any other constraints.

For simplicity, suppose that A, A^c, B, B^c each has positive lower probability, so independence of A and B depends only on unconditional previsions. Call \underline{P} an **independent lower envelope** when it is a lower envelope of some class \mathcal{M} of linear revisions such that A and B are independent under each linear revision in \mathcal{M} .⁸ A sensitivity analyst would regard A and B as independent only when his model \underline{P} is an independent lower envelope.

It is clear from Definition 9.1.1 that if \underline{P} is an independent lower envelope

9.1 INDEPENDENT EVENTS

then A and B are independent under \underline{P} . Consider the converse: if A and B are independent under \underline{P} , must \underline{P} be an independent lower envelope?⁹ The next example shows that the answer is no, even if \underline{P} is restricted to the field of events generated by A and B .

9.1.6 Example

Let $\Omega = \{1, 2, 3, 4\}$, $A = \{1, 2\}$, $B = \{1, 3\}$, $C = \{1, 4\}$. Define \underline{P} to be the lower envelope of the six linear previsions which assign probabilities $(0.3, 0.3, 0.2, 0.2)$ to some permutation of $(1, 2, 3, 4)$. This generates unique conditional probabilities which satisfy

$$\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A) = 0.4, \quad \bar{P}(A|B) = \bar{P}(A|B^c) = \bar{P}(A) = 0.6,$$

and similarly with A and B interchanged. Thus A and B are independent under \underline{P} . However, the values $\underline{P}(C) = 0.4$ and $\bar{P}(C) = 0.6$ are not achieved by any linear revision in $\mathcal{M}(\underline{P})$ under which A and B are independent; in fact these all satisfy $0.495 \leq P(C) \leq 0.505$. Thus \underline{P} is not an independent lower envelope.

The judgement that A and B are independent determines the probability of C quite precisely under a sensitivity analysis interpretation. Thus the constraints imposed by independence are stronger for a sensitivity analyst than for us. The effect is more pronounced when several judgements of independence are made. In the example, the three events A, B, C are pairwise independent under \underline{P} , but there is a unique linear revision in $\mathcal{M}(\underline{P})$ (the uniform distribution) under which they are pairwise independent.

We can modify this example as follows to show that many events can be pairwise independent under a coherent model \underline{P} , but not under any linear revision in $\mathcal{M}(\underline{P})$.

9.1.7 Example

Let $\Omega = \{1, 2, 3, 4, 5\}$ and $A_j = \{1, j\}$ for $2 \leq j \leq 5$. Suppose You judge the events A_2, A_3, A_4, A_5 to be pairwise independent and You also judge A_2 to be independent of $\{2, 3\}$. You also assess the upper probability of each singleton to be $\frac{1}{2}$. Let \underline{P} denote the natural extension of the quantitative assessments, so the extreme points of $\mathcal{M}(\underline{P})$ are the ten permutations of $(\frac{1}{2}, \frac{1}{2}, 0, 0, 0)$. If A and B are any two-point sets with one point in common then it is necessary for coherence with \underline{P} that $\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A) = 0$ and $\bar{P}(A|B) = \bar{P}(A|B^c) = \bar{P}(A) = 1$, so that A and B are independent. Thus the model \underline{P} is consistent with all the independence and probability judgements. It is easily verified that there is no linear revision that is consistent with all the judgements.¹⁰ The assessments are coherent and

apparently reasonable, but a sensitivity analyst would have to reject them as unreasonable.

Whereas models \underline{P} based only on direct judgements of desirability are compatible with a sensitivity analysis interpretation (see section 4.1), models based also on judgements of independence are not. The sensitivity analysis interpretation can be retained only by adopting a more restrictive definition of independence, which would go beyond the simple behavioural interpretation of Definition 9.1.1.

9.2 Independent experiments

It is perhaps more natural to regard independence as a property of experiments rather than events. Two experiments will be called (epistemically) independent if observing the outcome of either one does not change Your beliefs about the other. Let \mathcal{X}_1 and \mathcal{X}_2 denote the possibility spaces for the two experiments. We will assume that the experiments are logically independent, so that all pairs of outcomes (x_1, x_2) are possible and the joint possibility space Ω is the product space $\mathcal{X}_1 \times \mathcal{X}_2$. (That is intuitively necessary for the experiments to be epistemically independent.) We can then use the results of Chapters 7 and 8 in defining and establishing independence. The formal definition is as follows.

9.2.1 Definition

Suppose that $\Omega = \mathcal{X}_1 \times \mathcal{X}_2$, and let \mathcal{F} be a linear space of $\mathcal{C} \times \mathcal{S}$ -measurable gambles, as in assumptions 7.3.2. (Here \mathcal{C} is a σ -field of subsets of \mathcal{X}_1 , \mathcal{S} is a σ -field of subsets of \mathcal{X}_2 .) Suppose that \underline{P} , $\underline{P}(\cdot | \mathcal{X}_1)$ and $\underline{P}(\cdot | \mathcal{X}_2)$ are each defined on \mathcal{F} and are coherent. Let P_1 and P_2 denote the \mathcal{X}_1 - and \mathcal{X}_2 -marginals of \underline{P} . Say that the two marginals (or the two experiments) are **independent** under this model when $\underline{P}(Y | \mathcal{X}_2) = P_1(Y)$ for all \mathcal{C} -measurable gambles Y and $\underline{P}(Z | \mathcal{X}_1) = P_2(Z)$ for all \mathcal{S} -measurable gambles Z .¹

Thus the two experiments are independent for You when Your beliefs and behaviour concerning the first, described by Your previsions $P_1(Y)$ for \mathcal{C} -measurable gambles Y , would not change if You learned the outcome of the second experiment, and similarly Your beliefs about the second experiment would not change if You learned the outcome of the first. Independence again has a straightforward behavioural interpretation.

9.2.2 Standard definition of independence

Independence of experiments is a property of the model \underline{P} , $\underline{P}(\cdot | \mathcal{X}_1)$, $\underline{P}(\cdot | \mathcal{X}_2)$. Just as for events, independence may be regarded as a property of the unconditional joint prevision \underline{P} alone, provided all the outcomes in the

9.2 INDEPENDENT EXPERIMENTS

marginal spaces \mathcal{X}_1 and \mathcal{X}_2 have positive lower probability. (Then \underline{P} uniquely determines $\underline{P}(\cdot | \mathcal{X}_1)$ and $\underline{P}(\cdot | \mathcal{X}_2)$ through the generalized Bayes rule.) In that case independence reduces to the condition that $\underline{P}(\{x\}G(Y)) = 0$ for all \mathcal{C} -measurable Y and x in \mathcal{X}_2 , and for all \mathcal{S} -measurable Y and x in \mathcal{X}_1 . In general this condition is necessary but not sufficient for independence.²

When the unconditional prevision P is linear, independent marginals must satisfy the **factorization property** $P(YZ) = P_1(Y)P_2(Z)$ for all \mathcal{C} -measurable gambles Y and \mathcal{S} -measurable Z , since

$$P(YZ) = P(P(YZ | \mathcal{X}_1)) = P(YP(Z | \mathcal{X}_1)) = P(YP_2(Z)) = P_1(Y)P_2(Z).$$

This factorization property is sufficient for independence provided all outcomes in \mathcal{X}_1 and \mathcal{X}_2 have positive probability.³ Indeed, it is then sufficient that the joint probability P factorizes on rectangles, $P(A \times B) = P_1(A)P_2(B)$ for all A in \mathcal{C} and B in \mathcal{S} . The standard (measure-theoretic) approach defines marginals to be independent under P when P factorizes in this way.

Our definition of independence is therefore equivalent to the standard definition when the joint probabilities are additive and all possible outcomes have positive probability. The following examples show that in other cases our definition is strictly stronger.

9.2.3 Example⁴

We construct a model for two experiments which are certain to yield the same outcome in the unit interval $\mathcal{X}_1 = \mathcal{X}_2 = [0, 1]$, but which are independent according to the standard definition. Let $\mathcal{C} = \mathcal{S}$ be the σ -field generated by the singleton sets, i.e. containing all countable subsets of $[0, 1]$ and their complements. Let the conditional previsions $P(\cdot | x)$ be degenerate at (x, x) , for all x in \mathcal{X}_1 or \mathcal{X}_2 . This represents certainty that one experiment will have the same outcome as the other. Let the common outcome have any continuous distribution on $[0, 1]$, such as the uniform distribution defined by $P(Y) = \int_0^1 Y(x, x) dx$. Then $P(Y) = P(P(Y | \mathcal{X}_1)) = P(P(Y | \mathcal{X}_2))$, so the model P , $P(\cdot | \mathcal{X}_1)$, $P(\cdot | \mathcal{X}_2)$ is coherent.

Intuitively the experiments are not independent – with probability one, the outcome of one determines the outcome of the other. They are not independent according to Definition 9.2.1 because, if x_1 represents a specific outcome of the first experiment and B is the event that the second experiment has outcome x_1 , then $B \in \mathcal{S}$ and $P_2(B) = 0$ whereas $P(B | x_1) = 1$.

Nevertheless the experiments are independent according to the standard definition. If $A \in \mathcal{C}$ and $B \in \mathcal{S}$ then $P_1(A)$ and $P_2(B)$ take the values 0 or 1 according to whether A and B are countable or not. The factorization property $P(A \times B) = P_1(A)P_2(B)$ therefore holds.

9.2.4 Cantelli–Lévy paradox (Example 7.3.4)⁵

Here the marginal spaces Θ and \mathcal{X} are each the set of positive integers, and each marginal prevision is identical to Q , a ‘uniform distribution’ on the positive integers. (Q is a linear prevision but is not countably additive.) We saw in Example 7.3.4 that there is no coherent model under which the two marginals are independent. For if A is the event that $\theta \leq x$, independence implies that $P(A|\theta) = Q(\{\theta, \theta+1, \dots\}) = 1$ for all $\theta \in \Theta$ and $P(A|x) = Q(\{1, \dots, x\}) = 0$ for all $x \in \mathcal{X}$, so that $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are incoherent.

But there are joint previsions on the product space $\Theta \times \mathcal{X}$ which have both marginals identical to Q and satisfy the factorization property. One such joint prevision is defined by $P(W) = Q(P(W|\Theta))$, where all the conditional previsions $P(\cdot|\theta)$ are identical to Q . If the gamble Y depends only on θ and Z depends only on x , $P(YZ) = Q(P(YZ|\Theta)) = Q(YP(Z|\Theta)) = Q(YQ(Z)) = Q(Y)Q(Z)$. Thus the factorization property holds for P , which does have both marginals identical to Q . However, it seems unreasonable to call the marginals ‘independent’. Knowledge of θ does not change beliefs about x , but knowledge of x must change beliefs about θ . As we have seen, it is incoherent to take all the conditional previsions $P(\cdot|x)$ to be identical to the Θ -marginal Q .

If we define $\underline{P}(\cdot|\mathcal{X})$ to be the vacuous conditional prevision then P , $P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are coherent. Of course the marginals are not then independent; marginal beliefs about θ are precise, but after observing x the updated beliefs about θ are vacuous. The factorization property holds, but learning x changes Your beliefs about θ .

9.2.5 Borel’s paradox (Example 6.10.1)

Here the latitude θ and longitude ψ have marginal densities $f_1(\theta) = \frac{1}{2} \cos \theta$ and $f_2(\psi) = (2\pi)^{-1}$, and joint density $f(\theta, \psi) = f_1(\theta)f_2(\psi)$. Because the joint density factorizes, θ and ψ are independent under the standard definition. They need not be independent in the sense of Definition 9.2.1. Factorization of the joint density is necessary but not sufficient for 9.2.1, without the extra assumption that all conditional previsions have densities equal to the marginal densities.⁶

It is sufficient, for instance, to assume that conditional densities are defined (for all θ and ψ) through the continuous version of Bayes’ rule (Theorem 6.10.4). This extra assumption would be justified if You knew that each variable was measured imprecisely, with precision that does not depend on the value of the other variable. The lesson learned from Borel’s paradox was that, without these substantive assumptions, the conditional previsions are not determined by the joint density.

9.2 INDEPENDENT EXPERIMENTS

For example, it is coherent to take the previsions conditional on ψ to be vacuous for all possible values of ψ , or to be uniform distributions for countably many values of ψ . In these cases, learning the value of ψ can change the probabilities concerning θ , and we would not want to say that θ and ψ are ‘independent’.

These examples indicate that, when the outcomes of the marginal experiments have zero probability, independence cannot be characterized merely in terms of unconditional previsions through the factorization property.⁷ The reason is that unconditional previsions do not determine conditional previsions.

Our definitions of independence, 9.1.1 and 9.2.1, were chosen because of their simple behavioural interpretation, in terms of the irrelevance of one observation to beliefs about another. On the other hand, the standard (factorization) definition involves only unconditional previsions and has the advantage of mathematical simplicity. As in other cases, we regard simplicity of interpretation as more important than mathematical simplicity.

9.2.6 Conditional independence

Statistical experiments are often physically independent, but that does not justify a judgement of epistemic independence unless the sampling model is completely known. In general, physical independence justifies a judgement of conditional epistemic independence. The experiments are regarded as epistemically independent conditional on a parameter θ which indexes the possible sampling models. In practice, judgements of conditional independence are likely to be much more common than judgements of absolute independence.

Definition 9.2.1 can be extended in a straightforward way to define conditional independence, by regarding all the previsions involved as conditional on Θ .⁸ Formally we consider the product space $\Omega = \Theta \times \mathcal{X}_1 \times \mathcal{X}_2$. Say that the two experiments with sample spaces \mathcal{X}_1 and \mathcal{X}_2 are **independent conditional on Θ** when $\underline{P}(Y|(\theta, x_2)) = \underline{P}(Y|\theta)$ for all θ in Θ , x_2 in \mathcal{X}_2 , and gambles Y that are functions only of x_1 , and a similar condition holds with x_1 and x_2 interchanged. Provided all outcomes in \mathcal{X}_1 and \mathcal{X}_2 have positive lower probability conditional on θ , the definition simplifies as before to conditions of the form $\underline{P}(\{x_i\}G(Y|\theta)|\theta) = 0$. (Then conditional independence is a property of the sampling model $\underline{P}(\cdot|\Theta)$ alone.)

9.2.7 Many independent experiments

The definition of two independent experiments can also be extended to cover any finite number of experiments. Call n experiments **independent**

when Your unconditional prevision for each gamble Y that is a function only of some subset of the n outcomes agrees with Your prevision for Y conditional on the joint outcome of the other experiments. (Thus Your beliefs about any subset of the experiments are unchanged when You learn the outcomes of any other experiments.) Provided all outcomes of the marginal experiments have positive lower probability, independence can be defined in terms of the unconditional prevision \underline{P} alone, by requiring that $\underline{P}(AG(Y)) = 0$ whenever A is an event specifying the outcomes of particular experiments and Y is a gamble not depending on these outcomes. Conditional independence of many experiments can be defined in a similar way, by regarding all the previsions as conditional on Θ .

9.3 Constructing joint previsions from independent marginals

Qualitative judgements of independence are commonly used in constructing joint previsions. The simplest type of assessment strategy, considered in this section, involves assessing marginal previsions for each of two experiments, judging that the experiments are independent, and forming beliefs about the joint experiment by natural extension of these judgements.

9.3.1 Product previsions

Formally, suppose that You assess coherent lower previsions \underline{P}_1 , \underline{P}_2 to describe Your beliefs about each of two experiments with possibility spaces \mathcal{X}_1 , \mathcal{X}_2 . Assume that the experiments are logically independent, so that the joint possibility space is the product space $\Omega = \mathcal{X}_1 \times \mathcal{X}_2$. Suppose also that You judge the two marginal experiments to be independent. By Definition 9.2.1, Your conditional previsions must be defined by $\underline{P}(Y|\mathcal{X}_2) = \underline{P}_1(Y)$ and $\underline{P}(Z|\mathcal{X}_1) = \underline{P}_2(Z)$, for all gambles Y or Z depending only on x_1 or x_2 respectively. We will indicate the agreement between conditional and marginal previsions by writing $\underline{P}_1(\cdot|\mathcal{X}_2)$ and $\underline{P}_2(\cdot|\mathcal{X}_1)$ instead of $\underline{P}(\cdot|\mathcal{X}_2)$ and $\underline{P}(\cdot|\mathcal{X}_1)$. We will assume, as in Definition 9.2.1, that $\underline{P}_2(\cdot|\mathcal{X}_1)$ and $\underline{P}_1(\cdot|\mathcal{X}_2)$ are defined on a space \mathcal{F} of $\mathcal{C} \times \mathcal{S}$ -measurable gambles.

An unconditional lower prevision \underline{P} , defined on \mathcal{F} , will be called a **product prevision** or **independent joint prevision** with marginals \underline{P}_1 and \underline{P}_2 when \underline{P} , $\underline{P}_2(\cdot|\mathcal{X}_1)$ and $\underline{P}_1(\cdot|\mathcal{X}_2)$ are coherent. (It is necessary for coherence that \underline{P} does have marginals \underline{P}_1 and \underline{P}_2 .) A product prevision describes beliefs about the joint experiment in which the two marginal experiments are independent.

9.3.2 Compatible marginals

Given marginals \underline{P}_1 and \underline{P}_2 can always be extended to a joint prevision \underline{P} ,¹ but the Cantelli–Lévy paradox (Example 9.2.4) shows that \underline{P} cannot always

9.3 CONSTRUCTING JOINT PREVISIONS

be chosen to be an independent joint prevision. Call \underline{P}_1 and \underline{P}_2 **compatible marginals** when they can be extended to some independent joint prevision. There are two fundamental questions:

1. Which pairs of marginals are compatible?
2. What is the minimal product prevision with given (compatible) marginals?

These questions can be answered using the results of the previous chapters.

First, compatibility of the marginals \underline{P}_1 and \underline{P}_2 is equivalent (by Theorem 8.1.8) to coherence of the corresponding conditional previsions $\underline{P}_1(\cdot|\mathcal{X}_2)$ and $\underline{P}_2(\cdot|\mathcal{X}_1)$, which is equivalent (by Theorem 7.3.6) to a version of axiom S5.

Second, the minimal product prevision for two compatible marginals is simply the natural extension of the corresponding conditional previsions, which is defined by the formulas in Theorems 8.2.1 or 8.3.1.² It may be written, for X in \mathcal{F} , as $\underline{E}(X) = \sup \{\underline{P}_1(\underline{P}_2(Y|\mathcal{X}_1)): Y \in \mathcal{F}, \underline{P}_1(X - Y|\mathcal{X}_2) \geq 0\}$. By Lemma 8.3.2 we have the bounds $\underline{P}_1(\underline{P}_2(X|\mathcal{X}_1)) \leq \underline{E}(X) \leq \underline{P}_1(\bar{\underline{P}}_2(X|\mathcal{X}_1))$, so that $\underline{E}(X) = \underline{P}_1(\underline{P}_2(X|\mathcal{X}_1))$ when one marginal \underline{P}_2 is linear. Call \underline{E} the **independent natural extension** of \underline{P}_1 and \underline{P}_2 . It describes just those beliefs about the joint experiment that are entailed by the marginal assessments \underline{P}_1 and \underline{P}_2 and the judgement of independence.

When both marginals \underline{P}_1 and \underline{P}_2 are linear previsions, axiom S5 simplifies to S1 and compatibility is equivalent to the condition: there is no non-zero Y in \mathcal{F} with $\underline{P}_2(Y(x_1, \cdot)) > 0$ for all x_1 such that $Y(x_1, \cdot)$ is not identically zero, and $\underline{P}_1(Y(\cdot, x_2)) < 0$ for all x_2 such that $Y(\cdot, x_2)$ is not identically zero. Fubini's theorem can be used to show that this condition is satisfied whenever both \underline{P}_1 and \underline{P}_2 are countably additive. So any two countably additive marginals are compatible. The Cantelli–Lévy example shows that finitely additive marginals need not be compatible.³ This can be seen as another argument for countable additivity in the case of linear previsions, and against the model for a uniform distribution on the positive integers (section 2.9.5).

If the marginals are compatible and one of them (\underline{P}_2 , say) is linear then their independent natural extension is the unique product prevision with these marginals. That is because we can regard \underline{P}_1 as the \mathcal{X}_1 -marginal and $\underline{P}_2(\cdot|\mathcal{X}_1)$ as the linear prevision conditional on \mathcal{X}_1 , and these have a unique coherent extension $\underline{E}(X) = \underline{P}_1(\underline{P}_2(X|\mathcal{X}_1))$ by the results of section 6.7.

If the marginals are compatible and both linear, we obtain the formula⁴ $\underline{E}(X) = \underline{P}_1(\underline{P}_2(X|\mathcal{X}_1)) = \underline{P}_2(\underline{P}_1(X|\mathcal{X}_2))$, and the unique product prevision \underline{E} is also linear. In that case we will write $\underline{E} = \underline{P}_1 \times \underline{P}_2$ for the product prevision. When both marginals are also countably additive, their product prevision agrees (as a probability measure) with the product measure defined in measure theory. Although our definition of independence (9.2.1) is different

from the standard definition based on the factorization property, it does lead to the same product measure in the countably additive case.

9.3.3 Independent lower envelopes

Next we look more closely at the case where both marginals are non-linear. If they are compatible then they will have many different product previsions. How are these related to classes of linear product previsions, formed from pairs of dominating linear previsions? First, it is clear from Definition 9.2.1 that forming lower envelopes of models $\underline{P}_1, \underline{P}_2(\cdot|\mathcal{X}_1), \underline{P}_1(\cdot|\mathcal{X}_2)$ preserves independence of the marginals, and this also preserves coherence. It follows that lower envelopes of a class of pairs of compatible marginals are always compatible.⁵

For example, if \underline{P}_1 and \underline{P}_2 are each lower envelopes of countably additive linear previsions then they are compatible marginals. Let Γ index any class of countably additive pairs (P_1, P_2) such that $\{P_{\gamma}; \gamma \in \Gamma\}$ has lower envelope \underline{P}_1 and $\{P_{\gamma}; \gamma \in \Gamma\}$ has lower envelope \underline{P}_2 . Then each countably additive pair has a unique product prevision $P_1 \times P_2$. The lower envelope of these product previsions is a product prevision with marginals \underline{P}_1 and \underline{P}_2 , called the **independent lower envelope** of the class Γ . There may be many different classes Γ which give the same marginals $\underline{P}_1, \underline{P}_2$ as lower envelopes, and these give rise to different product previsions.⁶

It is natural for a sensitivity analyst to define the independent joint prevision generated by \underline{P}_1 and \underline{P}_2 to be some sort of independent lower envelope. But not all product previsions are independent lower envelopes. In particular, the independent natural extension of non-linear marginals is not an independent lower envelope, even in the simplest cases. That is illustrated by the next example.

9.3.4 Approximate Bernoulli trials

Consider two experiments each with two possible outcomes, $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$. These might represent the two possible ways in which a thumbtack can fall. Suppose that the two tosses of the tack are physically independent, but there is some indeterminacy in the tossing mechanism which produces imprecise chances. For simplicity, suppose that the chances on each trial j are known to be $\underline{P}_j(1) = 0.4$, $\bar{P}_j(1) = 0.5$. Because of their physical independence, the two trials are regarded as epistemically independent.⁷

Beliefs about the two outcomes will then be described by some product prevision with identical marginals \underline{P}_j . There are many such product previsions. In the absence of further information we would use the independent natural extension \underline{E} to describe beliefs about the joint experiment. This

9.3 CONSTRUCTING JOINT PREVISIONS

can be constructed as a lower envelope, using Theorem 8.1.10. It is the lower envelope of all additive joint probabilities P for which $P(\cdot|\mathcal{X}_1)$ dominates \underline{P}_2 and $P(\cdot|\mathcal{X}_2)$ dominates \underline{P}_1 , where the conditional probabilities are uniquely determined by P through Bayes' rule. By writing these conditions as constraints on P , we obtain eight inequalities involving the probabilities of joint outcomes, such as $3P(1, 1) \geq 2P(1, 0) \geq 2P(1, 1)$. The extreme points of $\mathcal{M}(\underline{E})$ can then be found by solving maximal sets of equalities. We find that $\mathcal{M}(\underline{E})$ has six extreme points, specified by their probability mass functions $P = (P(1, 1), P(1, 0), P(0, 1), P(0, 0))$: $P_1 = (0.25, 0.25, 0.25, 0.25)$, $P_2 = (0.16, 0.24, 0.24, 0.36)$, $P_3 = (0.2, 0.2, 0.3, 0.3)$, $P_4 = (0.2, 0.3, 0.2, 0.3)$, $P_5 = (\frac{2}{9}, \frac{2}{9}, \frac{2}{9}, \frac{3}{9})$, $P_6 = (\frac{2}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11})$. The independent natural extension \underline{E} is the lower envelope of these six linear previsions.⁸

Although \underline{E} was constructed as a lower envelope, it is not an independent lower envelope, i.e. it is not a lower envelope of linear product previsions. To see that, let A denote the event that the two trials have the same outcome. A linear product prevision which assigns probabilities α and β to the outcome 1 on the two trials has

$$P(A) = \alpha\beta + (1 - \alpha)(1 - \beta) = 2(\frac{1}{2} - \alpha)(\frac{1}{2} - \beta) + \frac{1}{2} \geq \frac{1}{2},$$

using the constraints $\alpha \leq \bar{P}(1) = \frac{1}{2}$ and $\beta \leq \frac{1}{2}$. But $\underline{E}(A) = P_6(A) = \frac{5}{11} = 0.455$, and this cannot be achieved by any linear product prevision whose marginals dominate each P_j .

Of the six extreme points in $\mathcal{M}(\underline{E})$, the first two are linear product previsions with identical marginals and the next two are product previsions with different marginals. But P_5 and P_6 , which achieve $\bar{E}(A)$ and $\underline{E}(A)$ respectively, are not product previsions, and consequently the independent natural extension is not an independent lower envelope.

Other product previsions with the same marginals can be constructed as independent lower envelopes from a subset of the linear product previsions $\{P_1, P_2, P_3, P_4\}$. For example, each of the subsets $\{P_1, P_2, P_3, P_4\}$, $\{P_1, P_2\}$, $\{P_3, P_4\}$, $\{P_1, P_2, P_3\}$ defines an independent lower envelope with marginals \underline{P}_1 and \underline{P}_2 .

9.3.5 Type-1 and type-2 products

Two types of independent lower envelope are especially natural under a sensitivity analysis interpretation. The **type-1 product** prevision with marginals \underline{P}_1 and \underline{P}_2 is defined as the lower envelope of all linear product previsions $P_1 \times P_2$ such that $P_1 \in \mathcal{M}(\underline{P}_1)$ and $P_2 \in \mathcal{M}(\underline{P}_2)$. This is the minimal independent lower envelope with marginals \underline{P}_1 and \underline{P}_2 . It describes beliefs about the joint experiment for a sensitivity analyst who regards the two experiments as independent and governed by linear previsions that are

known only to dominate the marginals \underline{P}_j , but which may differ between trials.

The type-1 product Q always dominates the independent natural extension \underline{E} with the same marginals. In the Bernoulli example, \underline{E} is the lower envelope of the six linear previsions P_1, \dots, P_6 , whereas Q is the lower envelope of the subset P_1, P_2, P_3, P_4 . For the event A that the two trials have the same outcome, Q is much more precise than \underline{E} : $Q(A) = 0.5$, $\bar{Q}(A) = 0.52$ whereas $\underline{E}(A) = 0.455$, $\bar{E}(A) = 0.556$.

However, the type-1 product must agree with the independent natural extension on rectangles $A \times B$, where A and B are subsets of \mathcal{X}_1 and \mathcal{X}_2 respectively. Then $Q(A \times B) = \underline{E}(A \times B) = \underline{P}_1(A)\underline{P}_2(B)$ and $\bar{Q}(A \times B) = \bar{E}(A \times B) = \bar{P}_1(A)\bar{P}_2(B)$.⁹

In the case where the marginals \underline{P}_1 and \underline{P}_2 are equal, as in the Bernoulli example, their type-2 product prevision is defined as the lower envelope of all linear product previsions $P \times P$ with identical marginals $P \in \mathcal{M}(\underline{P}_1)$. This model describes the beliefs of a sensitivity analyst who regards the two experiments as independent, identical, and governed by some linear prevision P known only to dominate \underline{P} .¹⁰

The type-2 product R dominates the type-1 product with the same marginals. In the Bernoulli example, let A denote 1 on the first trial and B denote 0 on the second trial. For the type-2 product $R(A \times B) = 0.24$ and $\bar{R}(A \times B) = 0.25$, whereas $Q(A \times B) = \underline{E}(A \times B) = \underline{P}_1(A)\underline{P}_2(B) = 0.2$ and $\bar{Q}(A \times B) = \bar{E}(A \times B) = 0.3$. Beliefs about the event $A \times B$ are much more precise under the type-2 product than under the independent natural extension and type-1 product.

9.3.6 Choice of product prevision

The Bernoulli example shows that, when both marginals are non-linear, the independent natural extension differs from the type-1 and type-2 products even in the simplest cases. Compatible non-linear marginals do not determine a unique product prevision.¹¹ Given only the marginal previsions and judgement of independence, the minimal product prevision \underline{E} , which describes the implications of these judgements, seems the appropriate model. Adopting this model does not rule out the possibility that Your beliefs can be described by a more precise product, such as the type-1 or type-2 product, but nor does it commit You to these models.

On the other hand, a sensitivity analyst would presumably model the same judgements through the type-1 or type-2 product, or possibly some other independent lower envelope.¹² There are problems in which it is natural to adopt these models, especially when the judgement of epistemic independence is based on knowledge of physical independence (as in section 9.6).

9.4 PERMUTABILITY

The type-1 and type-2 products are also easier to construct than the independent natural extension. But there is nothing in the original judgements (independence plus the marginal previsions) that entails these models. In the Bernoulli example, with A the event that the two outcomes are the same, You must be willing to accept an even-money bet on A if You adopt an independent lower envelope, but not if You adopt the independent natural extension \underline{E} . There is nothing in Your original judgements that implies such willingness. Again there is a clear difference between our direct interpretation, which models the judgements by \underline{E} , and a sensitivity analysis interpretation, which models the same judgements by the type-1 product or a more precise product.¹³

9.4 Permutability

We next examine two types of judgement that reflect symmetry in beliefs and evidence. Experiments are called permutable when Your previsions for gambles that depend on their outcomes are invariant under permutations of the outcomes. Permutability is a weak judgement which merely reflects indistinguishability of the experiments. A stronger judgement of exchangeability will be examined in the next section.¹

9.4.1 Definition

Suppose that n experiments have possibility spaces $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n$ that are formally identical. For any gamble X on the product space $\Omega = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ and any permutation π of the integers $(1, 2, \dots, n)$, define the **permuted gamble** πX by $\pi X(x_1, \dots, x_n) = X(x_{\pi(1)}, \dots, x_{\pi(n)})$. Suppose that a coherent joint prevision \underline{P} is defined on a subspace \mathcal{F} of $\mathcal{L}(\Omega)$, where \mathcal{F} is permutation-invariant in the sense that $X \in \mathcal{F}$ implies $\pi X \in \mathcal{F}$. Say that the experiments are **permutable** under \underline{P} when $\underline{P}(\pi X) = \underline{P}(X)$ for all gambles X in \mathcal{F} and all permutations π of $(1, 2, \dots, n)$.

As the experiments will be regarded as fixed in most of the following discussion, we will simply call the joint prevision \underline{P} permutable. We can define a permutation $\pi \underline{P}$ of \underline{P} by $\pi \underline{P}(X) = \underline{P}(\pi X)$. Then the condition for permutability of \underline{P} is simply that $\pi \underline{P} = \underline{P}$ for all permutations π , i.e. that \underline{P} is permutation-invariant.

This definition has a simple behavioural interpretation. You regard experiments as permutable when the desirability of gambles, and Your buying and selling prices for them, are unchanged by permutations of the outcomes on which the gambles depend. For example, the upper and lower probabilities You assign to a sequence of outcomes do not depend on the order of the outcomes.

Say that events A_1, \dots, A_n are **permutable** under \underline{P} when the corresponding binary experiments, with two possible outcomes representing occurrence or non-occurrence of A_j , are permutable. Thus Your upper and lower probabilities for A_j must not vary with j , and similarly for more complicated gambles that depend only on the events.

9.4.2 *Grounds for permutability judgements*

Permutability is not an arbitrary judgement. Like any other probabilistic judgement, it needs to be justified by reference to the available evidence.² The ground for a judgement that experiments are permutable is that the relevant evidence is symmetric in the experiments. In other words, any known differences between the experiments are irrelevant to beliefs about their outcomes. The judgements of relevance typically rely on information about causal connections, as in section 9.1.2.

Suppose, for example, that A_j represents the event that a particular person j lives to age 60. Whether You regard the events A_1, \dots, A_n as permutable should depend on Your information about the individuals. If the persons are distinguished only by their names, and listed in alphabetical order, then the events would usually be regarded as permutable on the grounds that they are ‘indistinguishable’, i.e. the available information that distinguishes the events is irrelevant to their occurrence. In other cases the current ages of the individuals might be known, or they might be ordered according to age, and that information would be relevant.

As a second example, let A_j denote the event that a thumbtack lands pin-up on toss j . In this case the events are ordered in time. If You believe that the tosses are physically identical, so that time is irrelevant, You should judge the events permutable. If You believe that the manner of tossing will change in time in a particular way then the ordering is relevant and the events are not permutable. Of course it is always possible that such changes will occur, but if You have no specific information about what changes are likely then Your evidence is order-invariant and You should judge the events permutable.³

Permutability is justified whenever the relevant evidence is order-invariant, and hence in cases of complete ignorance where there is no relevant evidence. Clearly the vacuous lower revision \underline{P} is permutable.

Judgements of permutability are similar in several respects to judgements of independence.⁴ Judgements of both kinds are usually made prior to any quantitative assessment of revisions, both are based on qualitative judgements of relevance, and both are useful in constructing revisions through natural extension.

9.4.3 *Permutable events*

Because the vacuous prevision is permutable, a judgement of permutability by itself does not produce a non-vacuous model. But when non-vacuous previsions are directly assessed, a permutability judgement has the effect of sharpening them. In the simple case of a sequence of n events, suppose that You assess precise probabilities p_j for the events that the sequence of outcomes consists of j successes followed by $n - j$ failures, for $1 \leq j \leq n$. This requires n precise assessments. Their natural extension to joint probabilities for the events is rather imprecise, e.g. the event that there are no successes in the sequence has upper probability $1 - \sum_{j=1}^n p_j$ and lower probability zero. But if You also judge the n events to be permutable then the probability of any sequence of outcomes is precisely determined. Permutability leads to a precise joint prevision. (Without the permutability judgement, You would need to make $2^n - n - 1$ further assessments of precise probabilities in order to construct a precise joint prevision.) Of course permutability will not produce a fully precise model in general, but it will increase the precision of the model.⁵

9.4.4 *Examples of permutable models*

A permutable joint prevision \underline{P} must have identical marginals \underline{P}_j . That is because, if X is a gamble in \mathcal{F} depending only on x_j , there is a permutation π with $\pi(j) = 1$ so that πX depends only on x_1 and $\underline{P}(X) = \underline{P}(\pi X)$, hence $\underline{P}_j = \underline{P}_1$.

When P is a linear joint prevision with independent and identical marginals P_j , P is uniquely determined by its marginals as the product prevision $P_1 \times P_2 \times \dots \times P_n$, and this is permutable. But when \underline{P} is non-linear it can have independent and identical marginals without being permutable.⁶ However, the three types of product prevision defined in section 9.3 do produce permutable joint previsions from identical marginals. In particular, the independent natural extension of identical marginals is always permutable.

It is clear from the definition of permutability that lower envelopes or convex combinations of permutable models are permutable. More generally, let θ index any class of permutable joint previsions $\underline{P}(\cdot | \theta)$ and let \underline{P}_0 be a lower prevision on $\mathcal{L}(\Theta)$, representing beliefs about the correct model. Then $\underline{P} = \underline{P}_0(\underline{P}(\cdot | \Theta))$ is a permutable joint prevision, since $\underline{P}(\pi X) = \underline{P}_0(\underline{P}(\pi X | \Theta)) = \underline{P}_0(\underline{P}(X | \Theta)) = \underline{P}(X)$. Hence we can construct permutable models from independent natural extensions or from type-1 or type-2 products, by forming lower envelopes or convex combinations or by assessing prior beliefs about the marginals. Examples will be given in sections 9.5 and 9.6.

9.4.5 Permutable classes of linear previsions

Next consider the relationship between permutability of a joint prevision and permutability of its dominating linear previsions. (Permutability coincides with the Bayesian notion of exchangeability in the case of linear previsions.) If all linear previsions in $\mathcal{M}(\underline{P})$ are permutable, their lower envelope \underline{P} must be permutable. But, as might be expected from the results concerning independence, permutable models \underline{P} can have dominating linear previsions that are not permutable.⁷

That is illustrated by the Bernoulli example 9.3.4. There the independent natural extension E is permutable, but it has dominating linear previsions P_3 and P_4 that are not permutable, as $P_3(1, 0) = P_4(0, 1) = 0.2$ but $P_3(0, 1) = P_4(1, 0) = 0.3$.

Here P_3 and P_4 can be obtained from each other by reversing the order of the two experiments, so $P_3 = \pi P_4$ and $P_4 = \pi P_3$ where π is the transposition of (1, 2). In general, a joint prevision \underline{P} is permutable if and only if $\mathcal{M}(\underline{P})$ is a **permutable class** of linear previsions, in the sense that when P is in $\mathcal{M}(\underline{P})$ so are all its permutations πP .⁸

How would a sensitivity analyst model permutation-invariance of evidence? He would presumably regard permutation-invariance as a constraint on the ‘ideal’ linear prevision. But permutability is a property of the class $\mathcal{M}(\underline{P})$ rather than of the individual linear previsions in it. The natural constraint for a sensitivity analyst is that \underline{P} should be the lower envelope of a class \mathcal{M} of linear previsions that are each permutable. Joint previsions of this form will be called exchangeable, and are studied in the next section. Exchangeable models are always permutable, but many permutable models, including independent natural extensions and type-1 product previsions, are not exchangeable. Once again our behavioural interpretation leads to models different from sensitivity analysis.

9.5 Exchangeability

We next define exchangeability. Our definitions of exchangeability and permutability are equivalent in the case of linear previsions, for which they agree with de Finetti’s concept of exchangeability.¹

A joint prevision \underline{P} will be called exchangeable when You are almost willing to exchange any gamble X for a permuted gamble πX . That is so when the gamble $\pi X - X$ has precise prevision zero under \underline{P} .²

9.5.1 Definition

A coherent joint prevision \underline{P} , defined on a permutation-invariant linear subspace \mathcal{F} of $\mathcal{L}(\mathcal{X}_1 \times \dots \times \mathcal{X}_n)$, is called **exchangeable** when $\underline{P}(\pi X - X) = 0$

9.5 EXCHANGEABILITY

for all X in \mathcal{F} and all permutations π of $(1, 2, \dots, n)$. In that case we will also say that the n experiments are exchangeable under \underline{P} . Events will be called exchangeable when the corresponding binary experiments are exchangeable.

It follows from this definition that if \underline{P} is exchangeable then $\bar{\underline{P}}(\pi X - X) = -\underline{P}(X - \pi X) = -\underline{P}(\pi^{-1}Y - Y) = 0$, where $Y = \pi X$. Thus all gambles $\pi X - X$ do have precise prevision zero.

Exchangeability implies permutability in general, since $0 = \bar{\underline{P}}(\pi X - X) \geq \underline{P}(\pi X) - \underline{P}(X) \geq \underline{P}(\pi X - X) = 0$ so that $\underline{P}(\pi X) = \underline{P}(X)$. When P is a linear joint prevision, $P(\pi X - X) = P(\pi X) - P(X)$ so that exchangeability coincides with permutability. Permutable non-linear models need not be exchangeable.³

It is clear from the definition that a lower envelope of exchangeable lower previsions is exchangeable.⁴ Hence \underline{P} is exchangeable provided all its dominating linear previsions are exchangeable. Conversely, an exchangeable \underline{P} has $\underline{P}(\pi X - X) = \bar{\underline{P}}(\pi X - X) = 0$, so that all dominating linear previsions satisfy $P(\pi X - X) = 0$ and are exchangeable. Thus exchangeability of a lower prevision is equivalent to exchangeability (permutability) of all its dominating linear previsions. For that reason, exchangeability rather than permutability is the appropriate reflection of symmetry for a sensitivity analyst.

9.5.2 Grounds for exchangeability judgements

Symmetry in the relevant evidence concerning experiments or events should be reflected in permutability of the lower prevision that models beliefs about them (section 9.4.2). The key question is: what further information is needed to justify the stronger judgement of exchangeability? Exchangeability means that all gambles of the form $\pi X - X$ have precise previsions. Substantial evidence is required to justify any precise assessment of previsions. To justify exchangeability we require information that the experiments are physically similar to each other, and not merely indistinguishable on present evidence. That requirement is supported by the following results, which imply that exchangeable events are analogous to repeated drawings from an urn (with or without replacement).

Some Bayesians have claimed that exchangeability is a weak assumption.⁵ But it has strong consequences. If You judge an infinite sequence of events to be exchangeable then You must be certain that the relative frequency of occurrence of the events will converge to a limit. That and other consequences of exchangeability indicate that it is actually a strong requirement. It seems a weak assumption to Bayesians because it is the only characterization of symmetry available to them; when previsions are assumed to be linear, exchangeability coincides with permutability. In cases where the

evidence is symmetric but there is little of it, neither linearity nor exchangeability is justified. That is clear when there is no relevant evidence at all. The evidence is then certainly symmetric. It can be properly represented by the vacuous joint prevision, which is permutable but not exchangeable. It is absurd to appeal to the symmetry in an absence of evidence to defend an exchangeable model, with its strong implications for precision of previsions and convergence of relative frequencies.⁶

The difference between permutable and exchangeable may be clarified by the two examples discussed in section 9.4.2. In the first, A_j represents the event that person j lives to age 60. If You are given no information about the individuals or how they were chosen, Your evidence is symmetric and You should judge the events permutable. But there is no reason to suppose that the people involved are similar in relevant characteristics (age, physical constitution, etc.), and exchangeability is not justified. Exchangeability would be justified, at least as an approximation, if You knew the people to be similar in those physical characteristics that have the greatest influence on propensities for living beyond 60.⁷

Secondly, consider repeated tosses of a thumbtack. A judgement of exchangeability will often be appropriate because You believe that the tosses are ‘repeated trials’ in the sense that they are performed in essentially the same way. That is, You believe that any differences between the tosses (such as time or precise location) do not influence their outcomes, so that the precise chances of outcomes do not vary between trials. You might judge the trials exchangeable even if You have no record of past tosses and Your beliefs about the operative chances are indeterminate. It is enough that You know that the tosses will be made under identical conditions. However, You should not regard the tosses as exchangeable if You suspect that the tossing method will vary in time. Often You will judge that some small degree of variation is likely, but have little idea about what form the variation will take. In that case permutable is justified but exchangeability is not. (An appropriate permutable model will be studied in section 9.6.)

9.5.3 Urn models

Suppose that events A_1, A_2, \dots, A_n are exchangeable under a linear prevision P . A fundamental result on exchangeability is that P can be interpreted as a model for drawings without replacement from an urn containing n balls. The balls are marked 0 or 1 and occurrence of A_j is identified with drawing a ball marked 1 on the j th trial. Each ball in the urn has an equal chance of being drawn, so the chance of any sequence of drawings is determined by the total number (m) of balls marked 1. The result is that the exchangeable linear previsions are just the urn models determined by some precise probability distribution for m .⁸

9.5 EXCHANGEABILITY

Now suppose that the events A_1, \dots, A_n are exchangeable under a lower prevision \underline{P} . Then \underline{P} is a lower envelope of exchangeable linear previsions. Hence \underline{P} can be interpreted as a model for drawings from the same urn, where now You have indeterminate beliefs about the composition m . Conditional on the number of events that occur (m), \underline{P} assigns the same precise probability to each sequence with exactly m ones.

So a judgement of exchangeability is effectively a judgement that the events are ‘like’ drawings from an urn. Exchangeability appears to be a rather strong judgement, requiring substantial information about the type of physical dependence between trials.⁹

9.5.4 Infinite exchangeable sequences

Now consider an unlimited sequence of trials, described by the infinite product space $\Omega = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots$ where each marginal space \mathcal{X}_j is $\{0, 1\}$. The occurrence of event A_j is represented by outcome 1 on trial j . Call the events exchangeable (or permutable) under \underline{P} when every finite sequence of events is exchangeable (or permutable) under \underline{P} .

The fundamental result concerning infinite exchangeable sequences is known as **de Finetti’s representation theorem**.¹⁰ This says that an infinite sequence of events is exchangeable under a linear prevision P if and only if P is a mixture of Bernoulli models. Formally, let $P(\cdot|\theta)$ denote the infinite product prevision with identical marginals $P(1|\theta) = \theta$. The exchangeable linear previsions are just those of the form $P(X) = P_0(P(X|\Theta))$ for some linear prevision P_0 on the Borel-measurable gambles of $\Theta = [0, 1]$.¹¹ For example, a specific sequence of n outcomes containing m ones has probability $\int_0^1 \theta^m (1-\theta)^{n-m} P_0(d\theta)$, which is a mixture of Bernoulli probabilities $\theta^m (1-\theta)^{n-m}$.

The linear prevision P_0 in this representation is essentially the limiting distribution of the relative frequency of ones in the sequence. The parameter θ can be interpreted as the limiting relative frequency of ones, or as the chance of 1 on each (physically independent) trial.¹² Then $P(\cdot|\Theta)$ is a standard Bernoulli sampling model, and P_0 represents prior beliefs about the unknown θ .

Suppose now that the events in the infinite sequence are exchangeable under a lower prevision \underline{P} . Exchangeability under \underline{P} is equivalent to exchangeability under every dominating linear prevision, by the similar result for finite sequences of events. It then follows from de Finetti’s theorem that a lower prevision \underline{P} is exchangeable if and only if it can be written as $\underline{P}(X) = P_0(P(X|\Theta))$, where $P(\cdot|\Theta)$ is the Bernoulli sampling model and P_0 is a coherent lower prevision on Borel-measurable gambles. Here P_0 represents Your indeterminate prior beliefs about Θ , and is simply the lower envelope of linear priors P_0 corresponding to P in $\mathcal{M}(\underline{P})$.¹³ The type-2

product previsions constructed from identical marginals are of this exchangeable form, with \underline{P}_0 representing the prior knowledge that θ is bounded by the marginal probabilities $\underline{P}(1)$ and $\bar{P}(1)$. Further examples of exchangeable models will be given in section 9.6.

Thus the exchangeable lower previsions on infinite sequences of trials are just the models for trials that are known to be physically independent and governed by identical chances.¹⁴ This result again shows that exchangeability is a strong property. It suggests that exchangeability is justified for infinite sequences only when the trials are known to be ‘repetitions’ of an experiment, in which the physical conditions do not vary between trials.

9.5.5 Convergence of relative frequencies

A strong consequence of exchangeability for infinite sequences is that it commits You to the belief that the relative frequency of ones after n trials, r_n , will converge to a limit as $n \rightarrow \infty$. Convergence cannot be settled in finitely many trials, so consider the observable event of ‘apparent convergence’.¹⁵ Say that the sequence of relative frequencies (r_n) **apparently converges** (j, k, δ) when the maximum and minimum of r_n over $j \leq n \leq k$ differ by less than δ , a small positive constant. In that case the relative frequencies are stable between trials j and k , at least to order δ .

For simplicity suppose that $j = \sqrt{k}$, regard δ as fixed,¹⁶ and let $C(k)$ denote the event that (r_n) apparently converges (\sqrt{k}, k, δ) . Then $P(C(k)|\theta) \rightarrow 1$ as $k \rightarrow \infty$ under each Bernoulli model $P(\cdot|\theta)$, and the convergence is uniform in θ . It follows that $\underline{P}(C(k)) \rightarrow 1$ as $k \rightarrow \infty$ for any exchangeable lower prevision \underline{P} . Provided You regard the infinite sequence of trials as exchangeable, for arbitrarily small δ there is some finite sequence length k such that You are practically certain that relative frequencies will apparently converge (\sqrt{k}, k, δ) . Similarly, after observing a sufficiently long sequence of outcomes You will be practically certain that the observed relative frequency will be reproduced arbitrarily closely in a long sequence of future trials.¹⁷

These consequences of exchangeability are disturbing if exchangeability is regarded merely as a reflection of symmetry in evidence. It seems absurd to suppose that when You have little or no information about an unlimited sequence of trials You must adopt strong beliefs about empirical matters such as the frequency behaviour of the sequence of outcomes. These consequences are not disturbing, however, if one recognizes that judgements of exchangeability must be based on substantial evidence that trials are physically independent and identical.

The empirical content of exchangeability is revealed also by the attitude

9.5 EXCHANGEABILITY

of Bayesians to exchangeable models. Suppose that a Bayesian judges an infinite sequence of events to be exchangeable. Conditional on any finite sequence of outcomes, future events should remain exchangeable for him. But in practice a Bayesian would abandon the exchangeable model if he observed that relative frequencies were unstable or that there was some pattern in the outcomes. That is difficult to understand from a personalist point of view, but it makes sense if one recognizes that the judgement of exchangeability is tantamount to the empirical hypothesis that trials are physically independent and identical. Any observations that are inconsistent with the empirical hypothesis should also lead You to abandon the exchangeable model.

As would be expected, a judgement of permutability does not commit You to a strong belief in convergence of relative frequencies. (Indeed, the vacuous joint prevision is permutable.) When \underline{P} is either an independent natural extension or a type-1 product constructed from identical non-linear marginals, so \underline{P} is permutable but not exchangeable, the upper and lower probabilities for apparent convergence $C(k)$ tend to one and zero respectively as $k \rightarrow \infty$, provided δ is smaller than the marginal imprecision of the model.¹⁸ Under these models Your beliefs about convergence of relative frequencies are vacuous. In general, any permutable model must have dominating linear previsions that are exchangeable, and it follows that the upper probability of apparent convergence must tend to one as $k \rightarrow \infty$. Thus You are not willing to bet against convergence of relative frequencies under a permutable model. You may or may not be willing to bet on convergence.¹⁹

9.5.6 De Finetti’s reductionism

De Finetti and other Bayesians have argued that aleatory probabilities can be eliminated from statistics by reinterpreting them in terms of exchangeable beliefs, using de Finetti’s representation theorem (section 9.5.4).²⁰ The theorem shows that exchangeable previsions for an infinite sequence of events are mathematically equivalent to Bernoulli models involving unknown chances. Bayesians can therefore use exchangeability to model sequences of experiments that would otherwise be modelled as Bernoulli trials, without needing to refer to chances or physical independence. (They may still refer to the Bernoulli parameter θ , but only as a convenient representation for exchangeable beliefs and not as an unknown chance.)

De Finetti gives several reasons for wanting to eliminate aleatory probabilities. First, his theory ascribes probabilities only to events that are observable, in the sense of section 2.11.9. Unknown chances, and other statistical parameters, are not observable.²¹ Second, he regards unknown

chances as ‘metaphysical’, ‘nonsensical’, ‘meaningless’ and ‘illusory’,²² apparently because he identifies ‘meaningful’ with ‘observable’.

Nevertheless, hypotheses about chances can be meaningful without being verifiable through direct observation.²³ Such hypotheses are not only meaningful, in our view, but also necessary in science. They are needed in order to model and explain the randomness that is observed in many natural phenomena. There would, of course, be advantages of economy in reducing the two concepts of probability (aleatory and epistemic) to a single concept, but it seems to us that both are needed.²⁴

The Bernoulli model for a sequence of events is, unlike de Finetti’s model, an empirical hypothesis. It asserts that the events are physically independent with identical chances. The Bernoulli model can explain features of the observed outcomes, such as apparent convergence of relative frequencies, in terms of this physical structure. De Finetti’s model, which involves only epistemic probabilities, can be used to predict future observations but not to explain them. If science aims to give explanations as well as predictions, aleatory probabilities seem indispensable.²⁵

The fact that de Finetti’s model makes no claims about physical structure or aleatory probabilities is, for Bayesians, one of its attractions. Another is that exchangeability appears to Bayesians a rather weak judgement, requiring only symmetry of evidence. Contrary to the Bayesian view, we have argued that exchangeability is a strong judgement that is justified (in the case of infinite sequences) only by evidence that the trials are physically independent and identical, the same evidence that justifies the Bernoulli model. De Finetti’s theorem supports this view. Thus exchangeability requires ‘evidence of symmetry’ (in physical structure) rather than just ‘symmetry of evidence’. This reverses de Finetti’s position, because the empirical hypotheses that he claims to eliminate through exchangeability are needed to justify a judgement of exchangeability.

Another argument against the reduction proposed by de Finetti is that, in order to construct the exchangeable model he suggests, You must contemplate an infinite sequence of events. That is contrary to de Finetti’s own emphasis on observable events. Unfortunately, the property of exchangeability for finite sequences of events is too weak to have the intended effect; not all finitely exchangeable models are mixtures of Bernoulli models.²⁶

Even if You can envisage an infinite sequence of events, it seems simpler (both conceptually and practically) to assess a prior prevision for an unknown chance θ than to assess exchangeable previsions for all possible (arbitrarily long) sequences of observations, even though the two assessments are mathematically equivalent by de Finetti’s theorem. That is true even if the aim is to use previous observations to form predictive probabilities

concerning future observations; the parameter θ may still be a useful intermediary.²⁷

9.6 Robust Bernoulli models

In this section we describe a robust version of the standard Bernoulli model for repetitions of a binary experiment. This illustrates how statistical models can be constructed from judgements of independence, conditional independence, marginal and conditional previsions, and vacuous beliefs about nuisance parameters.

9.6.1 Standard Bernoulli model

Consider a sequence of repeated trials with outcomes x_1, x_2, \dots, x_n taking the values 0 or 1. The trials might be tosses of a thumbtack, measurement of whether the patients arriving at a doctor’s surgery have blood pressure exceeding a certain level, or any other repeatable binary experiments. The standard Bernoulli model for such trials is that they are physically independent and identical, with some precise chance θ for the outcome 1 on each trial. These assumptions determine a precise sampling model for the sequence of outcomes $x = (x_1, \dots, x_n)$ conditional on θ . This sampling model is combined with prior beliefs about θ to give precise predictive previsions for x and precise posterior previsions for θ after observing x .

This Bernoulli model is appropriate when different trials are known to have the same physical characteristics, e.g. the thumbtack is tossed in the same way each time, or the patients arriving at the surgery are drawn randomly from the same population. It is possible that the tossing method will vary in time as the tosser becomes bored, and that the type of patient will vary (with respect to blood pressure) according to the time of day or, in the long term, according to changes in the population of potential patients. If the type of variation is predictable then it may be possible to construct a more complicated time series model which incorporates time variation in parameters and dependence between observations. But the choice of a complicated model is often arbitrary, because little is known about how conditions vary in time or because the trials have some intrinsic indeterminacy. In these cases precise sampling models can be constructed, but they will involve many parameters, about which little is known and little can be inferred from the observations.

9.6.2 Robust Bernoulli models

Suppose that outcome 1 has precise chance $(1 - \delta)\theta + \delta\psi_j$ on trial j , where θ, δ and ψ_j are all in the interval $[0, 1]$. Here θ is an average or ‘regular’

chance, ψ_j , represents the deviation in chances on trial j , and the constant δ measures the degree of imprecision or instability in the trials. The chance of 1 on trial j is a convex combination of θ and ψ_j with weights $1 - \delta$ and δ .

One interpretation of this model is that each trial is ‘regular’ with chance $1 - \delta$, and otherwise irregular. Regular trials, such as tosses of a thumbtack in a standard way or measurements on individuals of a standard type, produce outcome 1 with constant chance θ .¹ On irregular trials, such as tosses made in a different way or on a different surface, or any trials involving measurement or recording errors, the effective chances ψ_j vary in an unpredictable way.

Alternatively, the upper and lower bounds $\bar{P}(1|\theta) = (1 - \delta)\theta + \delta$ and $\underline{P}(1|\theta) = (1 - \delta)\theta$ might be interpreted simply as bounds on precise chances which vary in time, without any distinction between regular and irregular trials. A third interpretation is that $\bar{P}(1|\theta)$ and $\underline{P}(1|\theta)$ are upper and lower chances (indexed by θ) whose imprecision represents physical indeterminacy that is inherent in each trial (see section 7.2.9). Note that the lower prevision $\underline{P}(\cdot|\theta)$ is simply a linear–vacuous mixture, where the linear part is the standard Bernoulli model $P(1|\theta) = \theta$ and the vacuous prevision has weight $\delta = \bar{P}(1|\theta) - \underline{P}(1|\theta)$.²

As in the standard Bernoulli model, the trials are assumed to be physically independent, and hence they are epistemically independent conditional on the unknown parameters θ and $\psi = (\psi_1, \dots, \psi_n)$. These assumptions determine a precise sampling model $\underline{P}(\cdot|\Theta, \Psi)$.

To determine predictive previsions for \underline{x} and posteriors for θ conditional on \underline{x} , You must specify prior beliefs about (θ, ψ) . We will assume that θ and ψ are epistemically independent and that You have no information about variations in the nuisance parameters ψ_j , so that Your marginal prior prevision for ψ is vacuous.³ The prior prevision for (θ, ψ) is taken to be the independent natural extension of the vacuous prior for ψ and some marginal prior P_0 for θ .⁴

Models of this form, with positive values of δ , will be called **robust Bernoulli models**. (The case $\delta = 0$ defines the standard Bernoulli model.) A robust Bernoulli model is fully determined by the prior prevision P_0 concerning θ , together with the value of δ , which measures the degree of imprecision or robustness of the model. Since the prior P_0 is required also to apply the standard model, the robust model requires only the single extra assessment δ .

9.6.3 Predictive probabilities

Conditional on θ alone, the joint prevision for \underline{x} is simply the type-1 product (section 9.3.5) of the identical linear–vacuous marginals $\underline{P}(\cdot|\theta)$.⁵ So the trials are epistemically independent conditional on θ . If \underline{x} is a sequence of n outcomes containing m ones, for example, then

9.6 ROBUST BERNOULLI MODELS

$$\underline{P}(\underline{x}|\theta) = \underline{P}(1|\theta)^m \underline{P}(0|\theta)^{n-m} = (1 - \delta)^m \theta^m (1 - \theta)^{n-m}$$

and

$$\bar{P}(\underline{x}|\theta) = \bar{P}(1|\theta)^m \bar{P}(0|\theta)^{n-m} = (1 - \delta)^m (\theta + \varepsilon)^m (1 - \theta + \varepsilon)^{n-m},$$

where $\varepsilon = \delta/(1 - \delta)$.

The predictive probabilities for the outcomes \underline{x} are obtained from the prior P_0 through $\underline{P}(\underline{x}) = P_0(\underline{P}(\underline{x}|\Theta))$ and $\bar{P}(\underline{x}) = \bar{P}_0(\bar{P}(\underline{x}|\Theta))$. When $\delta = 0.1$ and P_0 is the uniform prior, for example, the single observations 0 and 1 each have [lower, upper] probabilities [0.45, 0.55], the pairs 00 and 11 each have probabilities [0.27, 0.37], and the pairs 01 and 10 each have probabilities [0.135, 0.235]. (Compare these with the precise probabilities $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$ under a standard Bernoulli model with uniform prior.)

Because the robust Bernoulli model is symmetric in the marginal experiments, the predictive prevision for \underline{x} is permutable, whatever prior P_0 is chosen for θ . It is not exchangeable, however, because ψ can be chosen asymmetrically to define dominating linear previsions that are not exchangeable.⁶ Compare with the standard model, which gives an exchangeable predictive prevision for any prior P_0 .

Whereas exact identity of the trials is assumed in the standard model and is required for exchangeability, the robust model assumes only approximate identity. Exact identity and exchangeability will often be unrealistic assumptions. (They imply certainty that relative frequencies will converge to a limit.) Under the robust model we expect fluctuations in relative frequencies to be no larger than δ , but nothing more is required. The model is consistent with long-run instability of relative frequencies, and δ measures the extent of plausible fluctuations.

9.6.4 Posterior previsions

Inferences about the parameter θ after observing outcomes \underline{x} are made by computing posterior previsions concerning (θ, ψ) through the generalized Bayes rule, and eliminating ψ by minimizing with respect to it.⁷ (Of course the marginal posterior for ψ will be vacuous as the prior is vacuous, so You can learn nothing about ψ under this model.) The inferences about θ can be partially described through the posterior upper and lower previsions for θ , which can be obtained from the prior P_0 together with the likelihood functions $L_1(\theta) = \theta^m (1 - \theta + \varepsilon)^{n-m}$ and $L_2(\theta) = (\theta + \varepsilon)^m (1 - \theta)^{n-m}$ respectively.

9.6.5 Example

The behaviour of these posterior upper and lower previsions for two different priors P_0 (one precise, the other imprecise) is shown in the table below. The

precise prior is the uniform distribution, which has constant density on $\Theta = [0, 1]$. Under this, θ has precise prior prevision 0.5. The imprecise prior is the lower envelope of the class of beta $(4, t)$ distributions over $0.25 < t < 0.75$, which have densities proportional to $\theta^{2c}(1-\theta)^{2(1-c)}$ for $0 < c < 1$. This assigns prior upper and lower previsions 0.75 and 0.25 for θ .⁸

Let I denote the identity gamble $I(\theta) = \theta$. The table gives the posterior upper and lower previsions $\bar{P}(I|\underline{x})$ and $\underline{P}(I|\underline{x})$ arising from outcomes \underline{x} with relative frequency $m/n = 0.2$ and various sample sizes n , for the following four models:

- (A) standard Bernoulli model and uniform prior
- (B) standard Bernoulli model and imprecise prior
- (C) robust Bernoulli model with $\delta = 0.1$ and uniform prior
- (D) robust Bernoulli model with $\delta = 0.1$ and imprecise prior.

Also shown are the posterior upper and lower predictive probabilities for outcome 1 on the next trial. These are given by

$$\underline{P}(x_{n+1} = 1|\underline{x}) = (1 - \delta)\underline{P}(I|\underline{x}) \quad \text{and} \quad \bar{P}(x_{n+1} = 1|\underline{x}) = (1 - \delta)\bar{P}(I|\underline{x}) + \delta,$$

because future trials are independent of the past conditional on θ . Under the standard Bernoulli models A and B, δ is zero and the predictive probabilities agree with the posterior upper and lower previsions of I , in the first three rows of the table.

Table. Posterior upper and lower previsions for $I(\theta) = \theta$ under models A, B, C, D, and posterior upper and lower predictive probabilities for outcome 1 on the next trial under models C, D. (All assuming observed relative frequency of ones is 0.2.)

		sample size (n)								
		0	5	10	15	20	25	50	100	1000
posterior previsions	A	0.50	0.29	0.25	0.24	0.23	0.22	0.21	0.206	0.201
	B	0.75	0.44	0.36	0.32	0.29	0.28	0.24	0.221	0.202
		0.25	0.22	0.21	0.21	0.21	0.21	0.20	0.202	0.200
	C	0.50	0.32	0.28	0.26	0.25	0.25	0.24	0.229	0.223
predictive probabilities	C	0.50	0.24	0.19	0.16	0.15	0.14	0.12	0.118	0.112
	D	0.75	0.49	0.40	0.35	0.32	0.31	0.27	0.246	0.225
		0.25	0.18	0.15	0.14	0.13	0.13	0.12	0.113	0.111
	C	0.55	0.38	0.35	0.33	0.33	0.32	0.31	0.306	0.301
D	C	0.45	0.21	0.17	0.15	0.14	0.13	0.11	0.106	0.101
	D	0.78	0.54	0.46	0.42	0.39	0.38	0.34	0.321	0.302
		0.22	0.16	0.14	0.13	0.12	0.11	0.11	0.102	0.100

Model A is precise and gives precise posterior previsions for the identity gamble I . Model B has imprecise prior but precise sampling model, and the posterior previsions approach precision as the sample size increases. For both A and B, the posterior distributions for θ converge to the degenerate distribution at 0.2 as the sample size increases: You are eventually confident that θ lies in an arbitrarily small interval around the observed relative frequency.

Similarly the posteriors for models C and D, based on the robust Bernoulli model, agree asymptotically. But the posterior upper and lower previsions for I tend to $\frac{2}{9}$ and $\frac{1}{9}$ respectively under the robust model. There are limits to the precision with which You can estimate θ , irrespective of how many observations are taken.

Under the robust model there are limits also to the precision of Your predictive probabilities concerning future trials. Even if θ were known precisely there would be imprecision δ in beliefs about the next trial, and the imprecision is larger than this because of the imprecise posterior for θ . The upper and lower probabilities for outcome 1 on the next trial tend to 0.3 and 0.1 under the robust model.

9.6.6 Asymptotic precision

The asymptotic precision of posterior beliefs under the robust model depends essentially on the degree of marginal imprecision, δ . For relative frequency $r = m/n$ and large sample sizes n , the posterior upper and lower previsions $\bar{P}(I|\underline{x})$ and $\underline{P}(I|\underline{x})$ are approximately equal to $r/(1 - \delta)$ and $(r - \delta)/(1 - \delta)$ respectively.⁹ The first value $r/(1 - \delta)$ is simply the natural estimate of θ assuming that the chance of 1 on each trial is the lower probability $(1 - \delta)\theta$, while the second value is obtained by assuming the chance is the upper probability $(1 - \delta)\theta + \delta$. After observing a large sample, the upper and lower predictive probabilities for outcome 1 on the next trial are approximately $r + \delta$ and $r - \delta$. The posterior imprecision can be measured by $\bar{P}(I|\underline{x}) - \underline{P}(I|\underline{x})$, which converges to $\delta/(1 - \delta) = \varepsilon$ as sample size increases, or by $\bar{P}(x_{n+1} = 1|\underline{x}) - \underline{P}(x_{n+1} = 1|\underline{x})$, which converges to 2δ . These quantities measure the limits to posterior precision that are imposed by the imprecision of the marginals $P(\cdot|\theta)$.

In practice we would be hesitant about drawing highly precise conclusions about the Bernoulli parameter θ , no matter how large the sample size. The limited precision under the robust Bernoulli models seems realistic. The imprecision δ might be much smaller than 0.1 for relatively stable physical processes such as coin tossing, but might be larger than 0.1 for some unstable or poorly understood phenomena such as economic processes.

9.7 Structural judgements

We end this chapter by outlining a general method of assessing probabilities which admits structural judgements, such as independence and permutability, as well as the direct judgements of desirability considered in Chapter 4.

9.7.1 Definition of structural judgements

As in Chapter 4, let \mathcal{D} be a class of gambles that You judge to be almost desirable. These are **direct judgements** of desirability (section 4.1.1). Let \mathcal{E} denote the coherent class of desirable gambles that is adopted as a model for Your beliefs. A direct judgement can be regarded as a constraint on \mathcal{E} of the form ' $X \in \mathcal{E}$ '. (If only direct judgements are made, \mathcal{E} is the smallest coherent subset of \mathcal{L} that contains \mathcal{D} , and it can be constructed from \mathcal{D} by natural extension.)

A **structural judgement** is defined to be a constraint on \mathcal{E} of the form 'if $X \in \mathcal{E}$ then $Y \in \mathcal{E}$ ' for specified gambles X and Y , or a set of such constraints. Informally, a structural judgement is a hypothetical judgement that if You were willing to accept gamble X then You would be willing also to accept gamble Y .

The most important types of structural judgements are **equivalence judgements** of the form ' $X \in \mathcal{E}$ if and only if $Y \in \mathcal{E}$ ', which is written as ' $X \in \mathcal{E} \Leftrightarrow Y \in \mathcal{E}$ '. These include judgements of independence and permutability. Equivalence judgements are natural in problems where there is a type of structural equivalence between gambles X and Y , e.g. where Y is a permutation of X , or where $Y = BX$ for some event B that is 'unrelated' to the outcomes on which X depends. The structural equivalence supports equivalence with respect to desirability: if one gamble is desirable then so is the other.

Although equivalence judgements are especially important, it is useful to consider the wider class of structural judgements. This includes judgements of non-negative dependence and comparisons between lower revisions ((f) and (g) below), which are not equivalence judgements.

9.7.2 Types of structural judgement

Of the judgements considered in this chapter, only exchangeability is a direct judgement. Exchangeability (Definition 9.5.1) simply means that all gambles $\pi X - X$ are almost desirable, for all X in \mathcal{F} and permutations π . Independence and permutability are structural judgements but not direct judgements. That is why exchangeability has a sensitivity analysis interpretation but independence and permutability do not.

9.7 STRUCTURAL JUDGEMENTS

Some general types of structural judgement are listed below, together with their translation into constraints on the class \mathcal{E} . (Compare with the list of direct judgements in section 4.1.1.)

(a) Permutable experiments (9.4.1)

A judgement that experiments are permutable imposes a set of equivalence constraints ' $X \in \mathcal{E} \Leftrightarrow \pi X \in \mathcal{E}$ ' for all X in \mathcal{F} and all permutations π . Permutability means just that the class of desirable gambles \mathcal{E} is permutation-invariant.

(b) Independent experiments (9.2.1)

A judgement that two experiments are independent imposes a set of equivalence constraints ' $X \in \mathcal{E} \Leftrightarrow BX \in \mathcal{E}$ ', where B is an event in one experiment and X is a function only of the outcome of the other experiment.¹

(c) Conditional independence (9.2.6)

This gives similar constraints to (b), except that X is replaced by AX where A is any event in the conditioning partition.

(d) Irrelevant events (9.1.1)

A judgement that B is irrelevant to A imposes constraints ' $\lambda A + \mu \in \mathcal{E} \Leftrightarrow B(\lambda A + \mu) \in \mathcal{E} \Leftrightarrow B^c(\lambda A + \mu) \in \mathcal{E}$ ' for all real λ and μ .

(e) Independent events (9.1.1)

Independence of A and B gives two sets of constraints like (d).

(f) Non-negative relevance and dependence

Say that event B is **non-negatively relevant** to A when $\underline{P}(A|B) \geq \underline{P}(A) \geq \underline{P}(A|B^c)$ and $\bar{P}(A|B) \geq \bar{P}(A) \geq \bar{P}(A|B^c)$. If also A is non-negatively relevant to B , call A and B **non-negatively dependent**. These judgements impose structural constraints like 'if $A + \mu \in \mathcal{E}$ then $B(A + \mu) \in \mathcal{E}$ '.

(g) Comparison of lower revisions

A judgement that You are willing to pay at least as much for gamble Y as for gamble Z , that is $\underline{P}(Y) \geq \underline{P}(Z)$, gives 'if $Z + \mu \in \mathcal{E}$ then $Y + \mu \in \mathcal{E}$ ' for all real μ . (Similarly for comparisons of conditional lower revisions or upper revisions.)

(h) Equality of lower revisions

A judgement that You have the same buying price for Y as Z , $\underline{P}(Y) = \underline{P}(Z)$, gives ' $Y + \mu \in \mathcal{E} \Leftrightarrow Z + \mu \in \mathcal{E}$ ' for all real μ .

(i) *Identical marginals*

Identity of marginal previsions gives a set of constraints of form (h).

(j) *Direct judgements (4.1)*

These can be regarded as a special type of structural judgement, of the form ' $X \in \mathcal{E} \Leftrightarrow 1 \in \mathcal{E}$ '. Since any coherent class \mathcal{E} must contain the gamble 1, this has the effect of the direct judgement ' $X \in \mathcal{E}$ '.

(k) *Nondesirability and indeterminacy*

Similarly, the equivalence judgement ' $X \in \mathcal{E} \Leftrightarrow -X \in \mathcal{E}$ ' effectively requires that X is not in \mathcal{E} . This represents a judgement that X is **nondesirable**, meaning that You do not have a disposition to accept X . This is different from a direct judgement that X is **undesirable**, meaning that $-X$ is desirable and You have a disposition to give away X . Undesirability implies nondesirability, but it is possible that X is neither desirable nor undesirable. In that case both X and $-X$ are nondesirable, and X is said to be **indeterminate**.

(l) *Exact assessments of previsions*

Suppose that You assess a lower prevision $\underline{P}(X)$ for gamble X . We have interpreted this assessment as a direct judgement, that $X - \mu$ is desirable for $\mu < \underline{P}(X)$, i.e. that You are willing to pay up to $\underline{P}(X)$ for X . Under an exhaustive interpretation of \underline{P} , it implies also a structural judgements, that $X - \mu$ is nondesirable for $\mu > \underline{P}(X)$, i.e. that You have no disposition to pay any more than $\underline{P}(X)$ for X .

Structural judgements include direct judgements, by (j), but not all probabilistic judgements are structural. For instance, bounds on upper variances, such as ' $\bar{V}(X) \leq 3$ ', are not structural judgements.² However, most probabilistic judgements do seem to be structural judgements.

9.7.3 *Elicitation using structural judgements*

We now outline how the theory of elicitation in Chapter 4 can be extended to admit structural judgements. Suppose that beliefs are elicited through a set of direct judgements \mathcal{D} and a set of structural judgements \mathcal{J} . Here \mathcal{J} is a class of ordered pairs (X, Y) , representing the judgements 'if $X \in \mathcal{E}$ then $Y \in \mathcal{E}$ '. The aim is to construct a coherent model for beliefs, \mathcal{E} or \underline{P} , from the elicited judgements \mathcal{D} and \mathcal{J} .

Say that the judgements \mathcal{D} and \mathcal{J} are **consistent** if there is some coherent class of desirable gambles \mathcal{E} that contains \mathcal{D} and satisfies all the structural constraints in \mathcal{J} . Consistency is the natural rationality condition here. (It

generalizes the condition that \mathcal{D} alone should avoid sure loss.) If the judgements are not consistent then there is no sensible belief-state that can give rise to them. If they are consistent then they can arise from any coherent model \mathcal{E} that contains \mathcal{D} and satisfies all the constraints in \mathcal{J} . Call the smallest such \mathcal{E} the **natural extension** of \mathcal{D} and \mathcal{J} . This is just the intersection of all coherent classes that contain \mathcal{D} and satisfy the constraints in \mathcal{J} .³ The lower prevision \underline{P} that corresponds to \mathcal{E} is the minimal coherent lower prevision that is consistent with the judgements \mathcal{D} and \mathcal{J} . (Intersection of classes \mathcal{E} corresponds to forming lower envelopes of lower revisions. Both coherence and structural constraints are preserved under intersections or lower envelopes.)⁴

There are two basic problems in this approach to elicitation:

1. to give necessary and sufficient conditions for consistency of judgements \mathcal{D} and \mathcal{J} ;
2. to construct the natural extension \mathcal{E} and corresponding \underline{P} directly from \mathcal{D} and \mathcal{J} .

The two problems were solved for the case of direct judgements in sections 4.1 and 4.2. In section 9.3 we gave some results for the case where \mathcal{J} contains judgements that experiments are independent, and \mathcal{D} contains judgements concerning the marginal experiments.

Some other cases can be handled by applying the structural constraints to \mathcal{D} to generate a larger class \mathcal{D}^* , and using the method of section 4.1 to obtain the natural extension of \mathcal{D}^* .⁵ Consider, for example, judgements of permutability. Suppose that \mathcal{J} consists of all pairs $(X, \pi X)$, representing the judgement that marginal experiments are permutable, and \mathcal{D} is any set of direct judgements concerning the joint experiment. Then \mathcal{D}^* , obtained by applying \mathcal{J} to \mathcal{D} , is the class of all permutations of gambles in \mathcal{D} . Now any coherent class \mathcal{E} that is permutation-invariant and contains \mathcal{D} must also contain \mathcal{D}^* . Hence it is necessary for consistency of \mathcal{D} and \mathcal{J} that \mathcal{D}^* avoids sure loss. If it does so, it has a coherent natural extension $\mathcal{E}(\mathcal{D}^*)$ that is permutation-invariant (because \mathcal{D}^* is), and $\mathcal{E}(\mathcal{D}^*)$ is the minimal coherent class that is permutation-invariant and contains \mathcal{D} . This shows that, in the case of permutability:

1. \mathcal{D} and \mathcal{J} are consistent if and only if \mathcal{D}^* avoids sure loss;
2. if so, the natural extension of \mathcal{D} and \mathcal{J} is just the natural extension of \mathcal{D}^* , which can be constructed using the methods of sections 4.1 and 4.2.

One way of constructing the natural extension $\mathcal{E}(\mathcal{D}^*)$ is through the corresponding class of linear previsions $\mathcal{M}(\mathcal{D}^*)$. In the case of permutability, $\mathcal{M}(\mathcal{D}^*)$ is the class of all linear previsions P such that $P(\pi X) \geq 0$ whenever X is in \mathcal{D} and π is a permutation. Then $\mathcal{M}(\mathcal{D}^*)$ is a permutable class of

linear previsions (section 9.4.5), although it may contain linear previsions that are not permutable.

9.7.4 Exhaustive interpretation

In earlier chapters we considered lower previsions \underline{P} based on direct judgements. Our interpretation of \underline{P} was that it represented what You had revealed about Your beliefs through the elicited judgements. This does not rule out the possibility that further judgements could be made to produce a more precise model. In particular, it does not rule out the possibility that ideal beliefs are determinate and representable by some linear prevision in $\mathcal{M}(\underline{P})$, as required by the dogma of ideal precision.

Matters are quite different when structural judgements are admitted. Suppose \underline{P} is the natural extension of both direct and structural judgements. Any model \underline{Q} that dominates \underline{P} is consistent with the direct judgements, but \underline{Q} need not be consistent with the structural judgements.⁶ Events or experiments which are independent (or permutable) under \underline{P} need not be independent (or permutable) under \underline{Q} . This has important implications for the sensitivity analysis and exhaustive interpretations discussed in section 2.10.

When \underline{P} is the unique model consistent with all the elicited judgements, it can be given an exhaustive interpretation. In principle, an exhaustive model can be constructed by making sufficiently many structural judgements. If, for example, You judge the gambles $A - \mu$ to be desirable for $\mu < 0.4$, undesirable for $\mu > 0.6$, and indeterminate (neither desirable nor undesirable) for $0.4 < \mu < 0.6$, then $\underline{P}(A) = 0.4$ and $\bar{P}(A) = 0.6$ are uniquely determined. Then \underline{P} will exhaustively model Your beliefs about A , although not necessarily about other gambles.

A judgement of indeterminacy requires more than mere uncertainty about whether You would accept $A - \mu$ if it was offered. It is a judgement that Your choice in such a situation is not determined by Your current state of mind. You have neither a disposition to accept the gamble nor a disposition to reject it.

You may often judge that many gambles are indeterminate, especially when You have little relevant evidence, but it is difficult to make sufficiently many judgements to obtain an exhaustive model. To do so You must, in effect, classify every gamble as desirable, undesirable or indeterminate.⁷

9.7.5 Sensitivity analysis interpretation

A lower prevision \underline{P} that is obtained by natural extension of a set of direct judgements has a sensitivity analysis interpretation, because the class of

dominating linear previsions $\mathcal{M}(\underline{P})$ is just the class of all linear previsions that are consistent with all the direct judgements. We have seen in this chapter that the sensitivity analysis interpretation of \underline{P} is inconsistent, in general, with structural judgements such as independence, permability and indeterminacy.⁸ There may be no linear previsions in $\mathcal{M}(\underline{P})$ which satisfy all the structural constraints that \underline{P} does.

A sensitivity analyst would (presumably) regard judgements of independence and permability as constraints on individual linear previsions, and translate them into models different from ours. (He would construct exchangeable models rather than permutable ones, for example.) Our approach has two obvious advantages.

First, we have defined independence and permability as structural properties of lower previsions. The simple behavioural interpretation of lower previsions carries over to the derived concepts. Now exchangeability, the sensitivity analyst's surrogate for permability, does have a behavioural interpretation, but we have argued that permability is the appropriate model for symmetry in beliefs. The implications of exchangeability, e.g. concerning convergence of relative frequencies, are unacceptably strong. It is less clear what behavioural interpretation can be given to a sensitivity-analysis concept of independence.⁹

Second, the sensitivity-analysis concepts are narrower than ours. Models which satisfy the sensitivity-analysis definitions of independence will satisfy our definitions, but not vice versa. (Similarly for exchangeability and permability.) Consequently we can model some sets of judgements that a sensitivity analyst cannot. He must dismiss the judgements as 'irrational', although they may be quite reasonable from our point of view (e.g. Example 9.1.7).

These are strong arguments against the sensitivity analysis interpretation and the dogma of ideal precision from which it derives its plausibility. Whether or not the arguments are regarded as conclusive, it is clear that the differences in interpretation have practical, as well as philosophical, significance. Whereas direct judgements generate the same model under both interpretations, sensitivity analysts will translate judgements of independence and permability into probability models that are different from ours, and the models can lead to different conclusions. Because structural judgements have a central role in practical statistics, we expect the difference in interpretation to lead to important differences in practice.

Notes

Chapter 1

Section 1.1

1. One can distinguish between strict Bayesians, who adhere to the dogma of precision, and Bayesian sensitivity analysts, who adhere to the weaker dogma of ideal precision (1.1.5). We will generally use the term ‘Bayesian’ to refer to strict Bayesians, such as de Finetti, Jeffreys, Savage and Lindley. Good (1965, p. 10) calls them extreme Bayesians. Bayesian sensitivity analysts are called ‘robust Bayesians’ by Berger (1984) and Levi (1985). Good (1965, p. 10) suggests that ‘one is *more or less* a Bayesian depending on the precision with which one is prepared to make intuitive probability estimates’, although Good (1971b) uses the term ‘Bayesian’ in a much wider sense. All previous authors who have called themselves Bayesians seem to have accepted at least the dogma of ideal precision, so it is proper to describe the present theory as non-Bayesian.
2. It seems appropriate to call this a dogma because it is a fundamental assumption of the Bayesian approach, it embodies a particular conception of rationality (Fine calls it a ‘myth of reason’), and it is accepted by many Bayesians without serious examination. The origin and history of the dogma of precision are discussed by Hacking (1975), Fine (1978), Walley and Fine (1979, pp. 325–9). Concerning the role of dogma in the development of science, see especially Kuhn (1962, 1963).
3. For Bayesians, probabilities are precise whether they are based on a large sample of observations or on ignorance. In fact, a Bayesian prior probability may be unchanged by the observation of many tosses. This is the ‘paradox of ideal evidence’ discussed in 5.3.4. The importance of distinguishing well grounded probabilities from those based on ignorance has been recognized by many authors. (See 1.8.2 for references.)
4. Indeed, there may be no constructive method of assigning precise probabilities to all the other events. Consider Lebesgue measure, defined on the measurable subsets of the unit interval. This has extensions to finitely

additive probabilities defined on all subsets, but their existence can be established only through non-constructive results such as the Hahn–Banach or separating hyperplane theorems. (See 3.4.2.) Imprecise extensions can always be defined constructively, by natural extension.

5. This is discussed in 6.9. Indeterminacy also arises when we condition a precise probability model on an event of probability zero (6.10).
6. There are various rules for combining the probability assessments of group members, such that the imprecision of group probabilities reflects the extent of disagreement amongst members. One simple rule of this kind is discussed in 5.2.3 and 4.3.9. Other rules are compared in Walley (1982).
7. Various ‘objective’ methods of statistical inference also produce precise posterior probabilities; see 7.4.7. For fundamental objections to frequentist methods of statistical inference, see 2.11.1, 5.8.4 and 7.5.
8. Robustness is characterized in this way by Dempster (1975) and Berger (1985a). Instead of constructing a ‘realistically wide class’ of precise models, as sensitivity analysts do, we would prefer to make realistically imprecise assumptions and assessments. The two approaches are compared by Walley and Pericchi (1988). For other concepts of robustness, see Huber (1981).
9. See 7.8.6 or 8.6 for a definition of the likelihood function.
10. For example, consider a class of Normal priors with mean μ and a range of variances. (The class defined in 7.8.3 is one possibility.) Allowing a range of prior variances from $\frac{1}{4}$ to 4 would produce an interval of posterior means for θ equal to [5.7, 6.3] in case (i), but [4.5, 7.5] in case (ii). More robust models are studied by Walley and Pericchi (1988). Models for prior–data conflict when the data consist of binary observations are examined in 5.4.
11. These departures are summarized in 2.10.5, and discussed more fully in sections 3.9, 5.2, 5.9, 6.9.8, 9.1, 9.4.5 and 9.7.5.

Section 1.3

1. We are concerned here with only a few simple distinctions. More thorough studies of interpretations of probability have been made by Fine (1973, 1983), Kyburg (1970) and Weatherford (1982).
2. This point is emphasized by Good (1950, 1952), Fine (1973), and de Finetti (1972, 1974, 1977b). De Finetti’s work is especially valuable as an example of how the interpretation can profoundly affect the mathematical theory. His emphasis on finite (rather than countable) additivity and on exchangeability are consequences of his operationalist interpretation.
3. The terms ‘epistemic’ and ‘aleatory’ are taken from Hacking (1975), who argues that the distinction can be traced back to 1660. A similar distinction was made by Poisson, Cournot, De Morgan and Venn in the last century,

and more recently by Carnap (1962), Mellor (1971), Levi (1980), and many other authors. The terms ‘empirical’, ‘physical’ and ‘objective’ are sometimes used in place of ‘aleatory’, and ‘subjective’ instead of ‘epistemic’. ‘Subjective’ and ‘objective’ are avoided because of their other connotations (see 2.11.2). In any case, epistemic probabilities can be ‘objective’, under a logical interpretation. Some further distinctions that are especially relevant to imprecise probabilities are made in 2.10.

4. A similar distinction can be made between frequentist concepts, in which probability models the behaviour of relative frequencies, and propensity concepts, in which probability models a relation between possible occurrences and underlying physical structure. A different distinction can be made between the behavioural and evidential interpretations of statistical procedures – see note 2 of section 1.5.

5. In quantum mechanics, for example, Schrödinger’s equation can be used to determine aleatory probabilities from other properties of the system (such as electrical potential). See Mellor (1971) for other examples.

6. It also seems clear that the epistemic probabilities used to measure uncertainty in conclusions should be posterior probabilities, based on all the available evidence. Compare with frequentist measures such as confidence coefficients, which are prior probabilities (1.5.3).

7. See Suppes (1984) for discussion of these examples. De Finetti (1937, 1975) has argued that aleatory probabilities can be eliminated from statistics, using ideas of exchangeability. His argument is criticized in 9.5.6.

8. The frequentist interpretations apparently date from Ellis and Cournot in the 1840s. They were developed by Venn (1888), von Mises (1957) and Reichenbach (1949). For discussion and criticisms, see especially Fine (1973, Ch. 4), Kyburg (1970, 1974a), Gillies (1973) and Salmon (1966).

9. Propensity interpretations have been proposed by Popper (1959b, 1983), Hacking (1965), Levi (1967, 1980), Mellor (1971) and Giere (1973, 1976). For criticism, see Kyburg (1974b).

10. A fuller statement is given in 7.2.4. Of course, this raises the question of how we can learn the values of aleatory probabilities. That is the main topic of Chapters 7 and 8.

Section 1.4

1. Throughout the book we refer to partial beliefs simply as ‘beliefs’.
2. For reviews of empirical evidence concerning the descriptive adequacy of the expected utility model, see Schoemaker (1982), von Winterfeldt and Edwards (1986), Edwards (1961), Slovic, Fischhoff and Lichtenstein (1977), Einhorn and Hogarth (1981), Slovic and Lichtenstein (1983), Pitz and Sachs (1984). The role of full (rather than partial) beliefs and values in explaining

human actions and evaluating their rationality is analysed in Taylor (1964), Norman (1971), and the essays in White (1968) and Raz (1978).

3. Dennett (1978, p. 272) and (1987, Ch. 1) argues persuasively that chess-playing computers, plants and even thermostats can be regarded (usefully) as intentional systems.

4. This convention is due to Good. It is also used by de Finetti (1974, 1975). ‘You’ has the additional virtue of being non-sexist. Nor does it discriminate against animals or machines. The term ‘we’ is used to refer to the various intentional systems, often conflicting or incoherent, who collaborated in writing the book.

5. Most behaviour is instinctive or spontaneous, but we are more concerned in this book with actions that result from conscious deliberation. The first account of partial belief as a behavioural disposition seems to be Ramsey (1926), for whom ‘the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it’. (Ramsey seems to have been influenced by the philosophy of pragmatism, especially by Peirce.) The idea that partial belief is a guide to action is much older (see Hacking, 1975), and Borel (1924) suggested that probabilities could be measured by observing their effects on betting behaviour. The dispositional interpretation of epistemic probability and utility is discussed more thoroughly by Jeffrey (1983) and Mellor (1971).

There is a large philosophical literature in which full belief is explained in terms of behavioural dispositions, e.g. Braithwaite (1946), Ryle (1949), Levi (1967), Armstrong (1973), and papers in Griffiths (1967) and Tuomela (1978). For general accounts of dispositions, see Mellor (1971, 1974) and Tuomela (1978).

6. Ryle (1949) calls them *many-track dispositions*.

7. Mellor (1980) calls them *quasi-dispositions* for that reason.

8. This is overlooked by personalist Bayesians, who assume that chosen betting rates or responses to scoring rules must reveal a person’s beliefs. See 5.6 and 5.7 for discussion. In fact, the Bayesian model is quite inconsistent with observed behaviour. (See the references in note 2.)

9. Dennett (1987) advocates a type of logical behaviourism. In experimental psychology, Tolman (1949) and Hull (1943) could be counted as logical behaviourists. (The term ‘neobehaviourism’ is used in psychology.) Tolman accepted theoretical concepts such as purposes, beliefs and values, which he called *intervening variables*, but he insisted that they should be ‘identified and defined in terms of the behaviors to which they lead’ (1949, p. 3). See Mackenzie (1977) and Smith (1986) for the history and philosophy of behaviourism.

10. There are alternatives to identifying beliefs with behavioural dispositions. Perhaps the most attractive of these is an evidential account, in

which beliefs are regarded as representations or summaries of a relation between hypothesis and evidence. (Evidential theories of probability are discussed in 1.5.4 and 1.7.3.) There are also functionalist accounts (Block, 1980), in which beliefs are defined by describing their causal relations with other beliefs and other mental states as well as their effects on behaviour; accounts such as Fodor (1975), Dretske (1981) or Armstrong (1973) in which beliefs are regarded as symbolic mental representations or semantic structures (like maps or pictures or sentences stored in the brain); and neurophysiological accounts in which beliefs are identified with particular states of the brain, e.g. with states of neural networks. These accounts have been influenced by recent work in cognitive psychology, reviewed in Hogarth (1975), Anderson (1980) and Gardner (1985), which has attempted to apply models of abstract information-processing systems to describe psychological states such as beliefs. Dennett (1987) discusses these alternative theories of belief. They all seem to be compatible with the minimal behavioural interpretation (1.4.4).

11. Watson (1930, Ch. 5) rejects behavioural dispositions. Skinner (1971, pp. 93–4, 118) rejects the concepts of beliefs, desires and purposes because they are not observable.

12. See especially de Finetti (1937, p. 148) and (1974, p. 145), and Lad (1990). De Finetti's two operational definitions are discussed in 5.7.1 and 5.7.6. For discussion of operational definitions in general, see 2.10.2 and Appendix H. Operationalism was advocated as a general philosophy of science by Bridgman (1927) and Skinner (1945). For criticisms, see Hempel (1965, 1966) and Mellor (1971).

13. Other objections to radical behaviourism and operationalism are that (a) introspection is a useful (but not infallible) source of evidence about beliefs and values (4.1); (b) observed behaviour is also a fallible source of evidence, due to arbitrary choice or failures of rationality; (c) many beliefs are never manifested in behaviour; (d) probabilities concerning unobservable states cannot be defined operationally, but can influence behaviour (2.11.9).

14. See 2.3.1 for a more specific statement.

15. Several extended interpretations are considered in 2.10.

Section 1.5

1. The distinction is discussed by Hacking (1965, Ch. 3), Levi (1967), Smith (1961), and various authors in *Synthese*, Vol. 36, No. 1 (1977). It seems more reasonable to regard decision as a special kind of inference, in which You conclude that a particular action is maximal or optimal. Levi (1967) argues that scientific inferences may usefully be regarded as decisions, but that the relevant values are 'cognitive' (rather than personal) values such as truth.

2. It is controversial whether, in the frequentist theory of statistics, the outcome of a hypothesis test should be taken to be a decision (to accept or reject the null hypothesis) or a conclusion (that the null hypothesis is consistent to a specified degree with the data). Neyman (1977) and most statistical textbooks adopt the first ('behavioural') interpretation, whereas Fisher (1956), Cox (1958), Kempthorne and Folks (1971), Cox and Hinkley (1974) and Birnbaum (1977) adopt versions of the second ('evidential') interpretation, which seems to be more useful and in closer accordance with most statistical practice. The corresponding interpretations of confidence intervals are discussed in 7.5.13.

3. This point was made by Smith (1961). The role in inquiry is emphasized by Levi (1980). Some objections to a behavioural theory of probability are discussed in 2.11.

4. According to the dispositional model (1.4.2), Your beliefs may be unconscious and unknown to You; see Mellor (1980). Experimental studies of introspection, such as those reported by Nisbett and Ross (1980, Ch. 9) or Valentine (1982, Ch. 5), suggest that it is difficult to elicit unconscious behavioural dispositions. On the other hand, there is no difficulty in eliciting probabilities and beliefs when these are constructed through conscious judgements.

5. This has been strongly emphasized by Shafer (1981a).

6. Compare these evidential theories with the behavioural theories of Ramsey, de Finetti, Savage and Jeffrey.

7. An example is given in 5.5.2.

8. Carnap (1971) and Shafer (1981a) clearly do accept the minimal behavioural interpretation, as well as an evidential interpretation. In his later writings, Carnap supported the principle of maximizing expected utility as the link between probabilities and behaviour, and suggested that his axioms might be justified by showing that they lead (via this principle) to reasonable behaviour. See 1.7 for further discussion of Carnap's theory, and the relationship between probability and evidence.

9. This question is quite independent of whether or not the probability-utility model in 1.4.1 is successful in explaining behaviour.

10. In some Bayesian theories, notably Savage (1972a), it is shown that every complete preference ordering which obeys certain rationality axioms can be obtained from a unique, precise probability measure and utility function. Bayesians have used such results to argue that preferences *should* be obtained from precise assessments of probabilities and utilities, but this argument is not convincing. First, it is necessary to assume that a preference ordering is defined on a set of 'acts' that is very much larger than the set of feasible actions. Second, the rationality axioms imposed are very strong, especially in requiring completeness. If the completeness axiom is dropped,

not all rational preference orderings can be constructed from separate assessments of imprecise probabilities and utilities, because of possible 'linkage' between probability and utility. (Example 7.8.5 illustrates an analogous type of linkage. See also Rubin (1985).) Finally, even if particular preferences could be constructed by assessing probabilities and utilities, that may not be the easiest way to construct them.

11. This is not to say that it is always easy to disentangle beliefs from values. Wishful thinking, in which beliefs are influenced by desires, is a common source of bias in decision making. See especially Elster (1979, 1983b), and Janis and Mann (1977).

12. A third activity of data description is more common than formal inference in statistical work. This involves summarization of data through graphical displays, descriptive statistics and models, but it need not involve assessments of probability. (See also 7.2.11.)

13. This approach is advocated by de Finetti (1974, Ch. 3.2) and Good (1950, 1983).

14. Both types of measurement require the availability of events whose chances are precisely known. Such events can be produced using coins, dice and lotteries.

15. Compare the simple probability axioms of de Finetti (1974) with the more complicated, decision-theoretic axioms of Savage (1972a).

Section 1.6

1. The term 'reflective thinking' is used by Dewey (1933).

2. The traditional notion of rationality, dating back to Aristotle, is concerned with choosing the best action to satisfy given desires, in the light of given beliefs. Aristotle calls this *practical wisdom* (*phronesis*). It is sometimes called means–end rationality or instrumental rationality. In Books 3 and 6 of the Nicomachean Ethics, Aristotle (1962) also emphasizes the need for 'moral excellence or virtue' as a constraint on desires: 'In an unqualified sense, that man is good at deliberating who, by reasoning, can aim at and hit the best thing attainable to man by action... virtue makes us aim at the right target, and practical wisdom makes us use the right means.' This and other conceptions of rationality are discussed by Rawls (1971), Kekes (1976), Simon (1978, 1982, 1983), Kyburg (1983b), Darwall (1983), Baron (1985), Nathanson (1985), Elster (1986) and Dewey (1933). The standard notion of rationality is criticized by Elster (1979, 1983b) and Pears (1984).

Bayesian decision theory generalizes the traditional account by allowing degrees of belief and desire, described through probabilities and utilities. Rationality involves choosing an action to maximize expected utility. (A

wider notion of rationality would apply also to the formulation of beliefs and desires.) For Bayesian views of rationality, see especially Good (1950, 1983), Savage (1972a), Carnap (1971), Jeffrey (1983), Vickers (1976), Eells (1982) and Suppes (1984).

3. Compare with the use of deductive logic in practical reasoning: internal rationality requires that You obey the axioms and follow the rules of logic, whereas external rationality requires that You select reasonable premises. Philosophers have made a similar distinction between concepts of rationality, e.g. Rawls (1971), Elster (1983b) and Darwall (1983).

Concepts of internal and external rationality apply also to utility models. These should satisfy axioms of coherence (internal rationality), but they should also be consistent with more deeply held values, and perhaps also with the values of a wider community and with principles of morality and justice; see Darwall (1983) or Rawls (1971). As suggested by Aristotle, some goals are more worthy than others. In any case, utility models need to be constructed through a process of reasoning and reflection about future benefits. Principles of morality and justice may play a role in this process that is similar to the principles of external rationality (1.7.2) in constructing probability models. It does appear, however, that principles of external rationality will usually impose greater constraints on probabilities than on utilities. (Utilities are typically 'more personal' than probabilities.)

4. The 'something You desire' can be taken to be a lottery ticket that has a small chance of winning a valuable prize (2.2). Your desires might themselves be harmful to You, but we could require that the prize is really beneficial to You, and not merely desirable as (say) a large quantity of cigarettes might be.

5. It is discussed by Savage (1972a), Tversky (1969) and Darwall (1983).

6. In fact, we assume only that You are willing to pay any price less than $P(A)$. This does not affect the argument.

7. The reverse inequality is not necessary for avoiding sure loss, but it is required in the Bayesian theory.

8. A similar view is expressed by Hill (1974, p. 557).

9. A loose way of expressing this conclusion is to say that, if You prefer X to Y and Y to Z , then You should prefer X to Z . This is not quite right, because it may be unreasonable to prefer X to Y and Y to Z . (These preferences may incur sure loss, or they may violate principles of external rationality.) A more correct statement is: it should not be the case that You prefer X to Y and Y to Z but do not prefer X to Z .

10. This is essentially the argument of Smith (1961, p. 7).

11. If the initial assessments $P(A) = 0.3$ and $P(B) = 0.4$ were precise, then $P(A \cup B) = 0.7$ would be their unique coherent extension. There are many problems in which precise assessments have a unique coherent extension but imprecise assessments do not, e.g. the extension of probabilities to

previsions (3.2), the extension of marginal and conditional previsions to joint previsions (6.7), or the extension of prior previsions and sampling models to posterior previsions (8.4, 8.5).

12. A strong argument against the adoption of axioms for purely mathematical reasons is given by de Finetti (1977b, pp. 381–5). He is especially critical of modern probability theory.

13. Of course, the axioms are tested also by examining their consequences. One also needs to question whether the interpretation is useful, and to compare it with alternatives. (The axioms might be justified under an interpretation that is entirely useless.) But we have argued that some kind of (minimal) behavioural interpretation is needed in probabilistic reasoning.

14. The distinction between finite additivity and countable additivity may be clarified by the examples in 2.4.5 and 6.8.5.

15. However, there are reasons for basing a theory on axioms of countable coherence. A type of countable coherence called full conglomerability is supported in section 6.8, and this rules out some models that satisfy the basic coherence axioms. It is not clear that all coherent, fully conglomerable models are reasonable, and we remain open-minded about whether further axioms can be justified.

Section 1.7

1. Some personalist Bayesians, notably Lindley (e.g. 1971a, 1983), have recognized this.

2. Loosely, evidence is relevant if taking it into account might change some of the probability assessments. A formal definition is given in 9.1.1, but judgements of relevance are typically made prior to the construction of a formal probability model. See 9.1.2 for discussion, and also Keynes (1921).

3. Kyburg (1974a) has suggested other principles that relate imprecise probabilities to imprecise knowledge of relative frequencies. One of these principles implies that, if You know only that a proportion r of the members of a class have some property, then You should adopt r as Your precise epistemic probability that a given member has the property. This principle seems too strong, unless it is known also that the given member is randomly chosen from the class.

4. Note also that application of the principles relies on judgements that certain evidence is irrelevant or that experiments are physically unrelated. The principles are not completely formalized.

5. Other theories, due to Jeffreys (1931, 1983) and Jaynes (1968, 1983), are criticized in 5.7, 5.12 and 5.5.

6. The problem is that Carnap implicitly relies on a principle of indifference, and this gives inconsistent probabilities when applied to different sets of predicates. (Although he denies that, in Carnap (1962, p. 518).) See 5.5.1 for further discussion.

7. The discussion in 5.5.1 indicates why. For details, see Fine (1973, p. 193). There are other, more technical, criticisms of Carnap's theory: (a) he assumes that the evidence can be completely specified in a simple, formal language, (b) he assumes that evidence is entirely observational (it cannot include theoretical models or assumptions), (c) he assumes a fixed language, and (d) universal generalizations about an infinite universe always have probability zero (on finitely many observations). For discussions of Carnap's theory, see especially Kyburg (1970), Ayer (1972) and Fine (1973, Ch. 7).

8. Levi (1980) argues for an 'objectivist inductive logic' that accepts the principle of direct inference as the only principle of external rationality.

9. That is the problem with the logical probabilities of Keynes (1921).

10. The idea of assessing probabilities indirectly, by calculating the implications of other assessments, is a central theme of Boole (1854, especially Ch. 18). It was also suggested by Good (1950, 1952, 1965), as a way of assessing probabilities or criticizing previous assessments, and by de Finetti (1937, p. 151). The step of widening the domain to include carrier events has been called 'extending the conversation to other events' by Lindley (1971a, 1983) and Tribus (1969).

11. These two aims are closely related, because intuitive judgements should be less precise than those based on a reliable model. Assessment strategies can therefore be regarded as ways of increasing the precision of the target probabilities. This is close to the view of de Finetti (1937, p. 151): 'The fact that a direct estimation is not always possible constitutes the reason for the utility of the logical rules of probability: their practical end is to relate an evaluation, itself not very directly accessible, to others by means of which the determination of the first evaluation is made easier and more precise.' For de Finetti, the discrepancy between the imprecision of probability assessments and the precision assumed in the Bayesian theory 'is only a consequence of that limited degree of idealization without which it would always be impossible to attain precision and to develop any theory at all'. But such 'idealization' prevents us from developing a theory to explain how some assessments can be 'more precise' than others.

12. The two strategies are compared in 9.5.6.

13. This type of model is called a hierarchical model. Such models can be very useful when there is sufficient evidence to assess the form of the sampling distribution for θ . See 5.10.4 for references. Empirical Bayes methods also attempt to model this kind of evidence; see Berger (1985a) for references. Mosteller and Wallace (1964) is an excellent study of various models and assessment strategies in a single, practical problem. (They consider several hierarchical models which have the same form as 4.)

14. This expertise is not easy to formalize or to teach, but see Cox and Snell (1981) or Mallows and Walley (1980) for discussion and examples. A case study in which expert opinions rather than statistical data were the

main source of information is described by Walley and Campello de Souza (1986).

15. See Edwards (1968), Slovic and Lichtenstein (1971), Hogarth (1975, 1980), Nisbett and Ross (1980), Cohen (1981), Kahneman, Slovic and Tversky (1982), Kyburg (1983b), von Winterfeldt and Edwards (1986, Ch. 13). Nisbett *et al.* (1983) describe some statistical heuristics that appear to be used successfully.

16. See the preceding references.

17. See Ramsey (1931, p. 199), Lichtenstein *et al.* (1977), Dawid (1982c), von Winterfeldt and Edwards (1986, section 4.6).

18. This is not meant to suggest that convention and tradition are adequate justifications for the use of particular strategies. Putnam (1981) argues that standards of rationality are not value-free.

19. On the role of context, see Levi (1980, Chs. 13, 18). On its role in statistical analysis, see Mallows and Walley (1980).

20. A general elicitation procedure is described in Chapter 4. (See especially note 1 of section 4.3.)

21. March (1978) is a good survey of this literature. Many of the ideas are due to Simon (1955, 1982, 1983). The discussion of ‘type-2 rationality’ in Good (1950) is also relevant. Sensitivity analysts such as Berger (1984) have suggested that it is only because of bounded rationality that we cannot live up to Bayesian ideals; see 5.9.

22. To allow for this, one could weaken the coherence conditions to apply to only a small number of assessments. Such conditions are defined in Appendix B.

23. If, for example, You assess an unconditional probability model, You should be able to extend this by defining conditional probabilities that are coherent with the unconditional ones. This is a substantive requirement, because there are some (‘non-conglomerable’) models that cannot be extended in this way (see 6.8).

24. The same procedure could be used as an assessment strategy for constructing prior probabilities.

25. It can be used to evaluate decisions in a similar way. To justify a decision, You should be able to make coherent assessments of probabilities and utilities that are consistent with evidence and any deeply held values, with respect to which the decision is maximal.

Section 1.8

1. The introduction to Kyburg and Smokler (1980) outlines the history of the subjective Bayesian theory. See also Fisher (1956, Ch. 2) and Lad (1990).
2. Hartigan (1983) and de Finetti (1972) present some more advanced mathematical theory. Lad (1990) is a textbook presentation of de Finetti’s approach, with some illuminating historical and philosophical discussion.

Pearl (1988) is a stimulating account from the point of view of artificial intelligence, with emphasis on computational methods and graphical representations. Cox and Hinkley (1974) and Barnett (1982) compare the Bayesian approach with other theories of statistical inference.

3. See especially his Chapter 19. Boole’s method seems to produce incorrect answers in some of his examples. See Hailperin (1976) for a detailed explanation of the method. According to Hacking (1975, Ch. 16) and Shafer (1978), James Bernoulli (1713) also admitted imprecise epistemic probabilities.

4. For other references on sensitivity analysis, see 5.9. Sensitivity analysis seems to be based on the dogma of ideal precision, that there are ideal, precise probabilities about which You are uncertain. It is then natural to model the uncertainty through second-order probabilities. That approach is discussed in 5.10. Nau (1986) uses ‘confidence weights’ rather than probabilities to model second-order beliefs.

5. It is noteworthy that this theory of imprecise probability predates both Kolmogorov’s (1933) theory of countably additive probability and the subjective Bayesian theories of probability. (A subjective theory was suggested by Borel (1924) and Ramsey (1926), in response to Keynes.) Although Keynes’s book was not published until 1921, almost all of it was written in the period 1906–1912.

6. This relies on a version of the principle of indifference, although Keynes (Ch. 4) rejects many of the classical applications of this principle.

7. Keynes’s view of probability appears to have influenced his theory of economics. In his theory, agents must act, but there are problems where they cannot base their actions on probabilities because ‘no solid basis exists for reasonable calculations’. That is, they may have little information about future events, or they may be unable to intuit the relevant probability relations. Knight (1933) takes a similar view.

8. Fine (1973, 1977a) and Fishburn (1986) survey this literature. De Finetti’s axioms are discussed in Appendix B5.

9. Theorem 3.3.3 is a general version of Smith’s result. In private communications, both Smith and Williams have made it clear that they did not intend a sensitivity analysis interpretation. The discussants of Smith (1961) do seem to regard his results as a justification for Bayesian sensitivity analysis. See de Finetti and Savage (1962) for further discussion of Smith’s paper.

10. For other criticisms of the Dempster–Shafer theory, see Williams (1978) and Walley (1987). Beran (1971) describes a distribution-free approach to statistical inference using belief functions, along the lines of Dempster’s approach. Suppes and Zanotti (1977), following Dempster (1967a), study a method for generating belief functions through multivalued mappings or ‘random relations’.

11. A similar theory of imprecise utilities can be obtained by dropping the

completeness axiom from the standard theory, due to von Neumann and Morgenstern (1947). This was outlined by Smith (1961) and Aumann (1962). See also Fishburn (1975). Jeffrey (1983, Ch. 9) suggested axioms for a complete preference ordering of propositions. This can be represented by precise probability and utility functions, but (unlike the Ramsey–Savage approach) the probability measure is not unique.

12. Compare with Example 3.9.10.

13. For further discussion of the various models for uncertainty, see Szolovits and Pauker (1978), Spiegelhalter and Knill-Jones (1984), Stephanou and Sage (1987), Pearl (1988), Lauritzen and Spiegelhalter (1988), the papers in Kanal and Lemmer (1986), Gale (1986) and Smets *et al.* (1988), and the papers by Shafer, Lindley and Spiegelhalter in *Statistical Science*, Vol. 2, No. 1 (1987).

14. Buchanan and Shortliffe (1984) and Shafer (1987) discuss the similarity between Dempster's rule for combining evidence and the rule used in MYCIN. The justification of Dempster's rule is considered in 5.13.

15. Inferences are carried out by propagating simple probability constraints, but these constraints are weaker than those obtained by natural extension. A serious limitation of the system is that it cannot update probabilities by conditioning on new evidence, as 'new evidence' can only be regarded as a further constraint on a fixed set of unknown, precise probabilities. Other computational methods for combining assessments of upper and lower probabilities are described by Andrews (1989).

16. Cohen's work is reviewed by Schum (1979). It is not clear whether Baconian probabilities have the behavioural interpretation of lower probabilities, or how they should be used, in general, as a basis for decision.

17. There are similar models for imprecise utilities. (In particular, classes of precise utility functions, or partial preference orderings of randomized consequences.)

18. That does not mean the choice is unimportant. Our reasons for presenting the theory in terms of lower previsions, rather than the alternative models, are discussed in 3.8.6. Models 4 and 5 appear to be more general than lower previsions (Appendix F).

Chapter 2

Section 2.1

1. These issues are discussed in more detail by Levi (1980, Ch. 1), de Finetti (1974, Ch. 2) and (1975, Appendix 4), and in the philosophical literature, e.g. White (1975).

2. The elements ω are called 'states of affairs'. (Strictly, they are mathematical representations of states of affairs.) We do not require the states to be identified with propositions or sentences. Many accounts of belief, e.g.

Armstrong (1973) or Jeffrey (1983), take *propositions* to represent the object or content of a belief. The analysis of propositions is controversial (see Dennett (1987)), and it is not clear that a theory of probability needs to refer to them. (On some accounts, e.g. Jeffrey (1983), propositions are merely representations of sets of possible states of affairs, which are often identified with sets of 'possible worlds'.) Of course, *sentences* in ordinary or formal language may be needed to describe or define the state represented by ω (or by a proposition). Hacking (1967) argues that it is more natural to define personal probabilities on sentences rather than propositions, since You may not know that two sentences describe the same proposition. See also Fodor (1975) and Block (1980).

3. See especially the Appendix of de Finetti (1975).

4. It may even be logically undecidable whether they are epistemically possible (Hacking, 1967). De Finetti discusses these problems in (1974, 2.1.3) and (1975, 8.8.3 and Appendix 4). Apparent possibility is discussed by Vickers (1965) and by Hacking (1967), who calls it personal possibility and argues that it is the appropriate concept of possibility for a personalist theory of probability. The example in the next paragraph is discussed by Hacking (1967) and Savage (1967).

5. You can always construct an epistemically exhaustive space by including the event 'anything else happens', denoted by A . But it may be difficult to assess the probability of A , or the utilities associated with it, especially if You have little idea what kind of possibilities it includes. See Fine (1973, IIIB).

6. These ideas are explained in section 4.3.7 and in Chapter 6.

7. The problem of deciding how much detail to include in ω is discussed by Savage (1972, 2.5 and 5.5), Shafer (1976, 1981a) and Manski (1981).

8. De Finetti (1974, section 2.4) gives further reasons. Quantum mechanics, which emphasizes the measurements that can be made on a system rather than its underlying state, also supports an approach based on random quantities. Coherence conditions equivalent to those studied in this chapter have been formulated in terms of random quantities by Williams (1975a).

Section 2.2

1. The gambles involved in any operational procedure for measuring probabilities must also be bounded. Under compelling axioms for preferences, either a utility function is bounded or there is a countable mixture of consequences that has infinite utility. See Chernoff and Moses (1959, Appendix F2).

2. De Finetti (1974, section 3.2.5) suggests that this is true for small amounts of money. (It may fail if the amounts are too small.)

3. Ramsey (1926) suggested a method of measuring utilities in terms of

known probabilities. The modern theory of utility is due to von Neumann and Morgenstern (1947), and is described by Lindley (1971a), Jeffrey (1983), Berger (1985a), and von Winterfeldt and Edwards (1986).

4. This section is based on Smith (1961). Smith attributes the idea of probability currency to Savage (1954). Savage (1971) attributes the idea to Smith (1961). Leamer (1986) describes several ways of eliciting upper and lower probabilities as selling and buying prices for lottery tickets.

5. The theory could be presented in terms of integer-valued gambles, without assuming that lottery tickets are divisible, as the coherence axioms (Definitions 2.4.1, 2.5.1) involve only finite sums of gambles.

6. The key property of this model is that it involves only two ‘consequences’, winning the prize or not winning, whose utilities can be taken to be 1 and 0 precisely.

7. Several assumptions are implicit in the following argument:

- (a) Accepting a gamble cannot influence the state ω or the value to You of the prize;
- (b) Your only interest is in winning the prize; gambling, and winning lottery tickets, have no value to You except as a means to this end;
- (c) You would prefer to win the prize, whatever the state ω .

If (a) fails, the arguments for D2 and D3 break down because two gambles may differ in desirability although they have the same utility under each state ω . Assumption (b) is needed to prevent You from accepting unfavourable gambles just because You enjoy gambling.

Two stronger assumptions are needed to ensure that Your probabilities can be accurately elicited using lottery tickets. (They do not seem to be needed to justify axioms D0–D3.)

- (d) The value to You of the prize does not depend on the true state.
- (e) The desirability of a gamble does not depend on Your holding of lottery tickets.

It is clear that Your attitudes to gambles may be distorted when these assumptions fail. Concerning (d), see Savage (1971, section 3).

8. D2 reflects the arbitrariness in the choice of the unit of utility. A utility gain of X can be regarded as a gain of λX in different units.

9. This kind of argument has been used to justify the ‘extended sure-thing principle’ of decision theory. For objections, see Allais and Hagen (1979).

10. The desirability axioms are examined more closely in Chapter 3 and Appendix F. For other methods of measuring utilities, see von Winterfeldt and Edwards (1986, Chs. 7 and 8), Farquhar (1984), and other references therein. It is a problem for experimental psychologists to investigate whether people obey the rationality axioms D0–D3 when rewards are in probability currency, or in other commodities such as money.

Section 2.3

1. The usual route in presenting a theory of probability is to first define probability, and then define expectation in terms of probability. Instead, we start with prevision (expectation) and then treat probability as a special case. The main reason is that, without the usual assumption of additivity, probabilities do not determine previsions (2.7.3). In taking prevision rather than probability to be fundamental, we follow de Finetti (1972, 1974). The 17th-century text of Huygens was also based on the concept of expectation (see Hacking, 1975), as was Bayes (1763), Suppes (1984, appendix), and the excellent introductory text of Whittle (1970). We prefer de Finetti’s term ‘prevision’ rather than ‘expectation’, so that the latter can be used in the standard sense of a functional constructed from probability (3.2).

Compare with the approaches of Lebesgue and Daniell to constructing a theory of integration. Lebesgue first develops a theory of measure, and then defines the integral in terms of measure. Daniell assumes the integral is defined on a linear space, extends it to a larger domain by taking monotone limits and differences, and recovers measure theory as a special case by restricting the domain of the integral. For the theory of the Daniell integral, see Daniell (1917–18), Shilov and Gurevich (1977), Loomis (1953) or Royden (1963).

2. For example, You might find it easier to order events according to which is more probable, or to assess distribution functions or quantiles. (See Chapter 4 for many other forms of assessment.)

3. This definition generalizes de Finetti’s (1974) definition of linear previsions (see 2.3.6). Equivalent (but slightly more complicated) sets of axioms are given by Williams (1976) and Huber (1981, section 10.2).

4. Axiom D0 is implicit in the requirement that $P(X)$ is a finite real number. With this requirement, it is easy to verify that coherence implies $P(X) \leq \sup X$, which corresponds to D0. This requirement (or D0) is needed to ensure finiteness of $P(X)$, as the assignment $P(X) = \infty$ for all $X \in \mathcal{X}$ satisfies axioms P1–P3.

5. See Leamer (1986) for relevant models. Observations of market buying and selling prices are commonly used in the theory of economic demand to model ‘revealed preferences’.

Section 2.4

1. There is an analogy with deductive logic. The set of probability judgements represented by P is analogous to a set of propositions. Incurring sure loss is analogous to deducibility of a contradiction from the set of propositions (logical inconsistency), whereas incoherence is analogous to a

failure to deduce all their logical implications. (The procedure of natural extension corresponds to deduction of all these implications.)

In the case of precise probabilities, the formal definition of avoiding sure loss is due to de Finetti (1931) and Lehman (1955). The idea was suggested earlier by Ramsey (1926), and it seems to be what Boole (1854, Ch. 19) meant by ‘statistical conditions’. For imprecise probabilities, Smith (1961) used the idea informally, Buehler (1976) gave an equivalent definition of ‘preference-reversal coherence’ in terms of partial preference orderings, and Huber (1981, p. 257) gave an equivalent definition for upper and lower probabilities.

2. We assumed only that $\underline{P}(X)$ is a supremum buying price for X . This implies that $G(X) + \delta$ is desirable for some $\delta < \varepsilon$. It then follows from D1 and D3 that $G(X) + \varepsilon$ is desirable.

3. In our terminology, an event is *sure* just when the complementary event is not possible. So the only sure event, with respect to the possibility space Ω , is Ω itself. It would be natural to extend this by calling an event A sure whenever $\underline{P}(A) = 1$. It is easily verified that You avoid sure loss in this extended sense, meaning that $\sup\{\sum_{j=1}^n G(X_j)(\omega): \omega \in A\} \geq 0$ whenever $X_j \in \mathcal{K}$ and $\underline{P}(A) = 1$, if and only if You avoid sure loss in the sense of 2.4.1. So avoiding sure loss has the same meaning for an epistemic possibility space Ω as for a (smaller) practical possibility space A (2.1.2).

4. This condition has an interesting geometrical interpretation. Regarding the possible reward vectors $\underline{X}(\omega) = (X_1(\omega), \dots, X_n(\omega))$ as points in \mathbb{R}^n , the condition fails when there is a hyperplane $\{\underline{X}: \sum_{j=1}^n \lambda_j X_j = \mu\}$, with each λ_j non-negative, which strongly separates the point $\underline{P}(\underline{X}) = (\underline{P}(X_1), \dots, \underline{P}(X_n))$ from the set of possible reward vectors. So the lower previsions \underline{P} defined on domain $\{X_1, \dots, X_n\}$ which avoid sure loss are just those whose points $\underline{P}(\underline{X})$ do not lie ‘above’ the set of possible reward vectors (in the sense that they can be separated from the set by such a hyperplane). De Finetti (1974, section 3.4) notes that the linear previsions correspond to those points which cannot be separated from the set of possible reward vectors by any hyperplane (not necessarily with $\lambda_j \geq 0$). These are just the points in the closed convex hull of the set of possible reward vectors. This shows that the set of lower previsions that avoid sure loss is much larger than the set of linear previsions.

5. The difference is clarified in 7.1.

6. For definitions of strict coherence, see Shimony (1955), Kemeny (1955) and Carnap (1971). The idea is to rule out finite sets of almost-desirable gambles $G(X)$ whose net outcome is surely non-positive and possibly negative. These conditions are not compelling because they rely on the assumption that $G(X)$ is really desirable. (Compare with the conditions in 7.1, which do not rely on that assumption.) Moreover, these conditions

of strict coherence imply that all non-empty subsets of Ω must have positive probability, which cannot happen when Ω is uncountable. A more general definition of strict coherence can be given in terms of a topology on Ω , and has the effect that all open subsets of Ω must have positive probability.

Countable coherence requires that certain countable sums of desirable gambles are not uniformly negative. (The sequences of gambles must satisfy some convergence restrictions.) See Heath and Sudderth (1972) and Walley (1984–85).

7. This is called pure or riskless arbitrage, to distinguish it from risk arbitrage, which involves some possibility of loss. Risk arbitrage has received considerable publicity from the efforts of Ivan Boesky and his associates. Boesky (1985, especially Ch. 8) is an edifying introduction to both kinds of arbitrage. Weisweiller (1986) is a good survey of arbitrage possibilities.

8. Strictly, α should be the minimum price for which the stocks can be bought in the market, when commission and other costs are included, which will be a little higher than the stock index. (Similarly, μ is the maximum selling price, after deducting costs.) The transactions are often executed in practice by computer trading programs and they may involve stock worth millions of dollars, relative to which the transaction costs are very low.

Section 2.5

1. The idea of coherence was used informally by Smith (1961, p. 8). Williams (1975a, 1976) gave a formal definition. Huber (1981, p. 259) gives an equivalent condition for the case of upper probability.
2. Coherence is much stronger than avoiding sure loss. For example, You can avoid sure loss by refusing to accept sure gains, e.g. assigning $P(1) = 0$. Compare with case 2.
3. If \mathcal{K} contains any constant gamble, we can require m and n to be positive integers in 2.5.1, and ignore cases 1 and 2.
4. This can be proved by modifying the proof of 2.5.5. Instead of Z in the proof, consider $\mu = \inf(X - Y)$, so that $X \geq Y + \mu$ and P4 applies.
5. P2 can be replaced by the weaker axiom P2' in the characterization of coherence on linear spaces (2.3.3, 2.5.5).

Section 2.6

1. This allows the possibility that $\underline{P}(X) = -\infty$. If all \underline{P}_i are coherent then $\underline{P}(X)$ is finite.
2. In fact, probability mixtures of arbitrarily many coherent lower previsions are coherent.
3. This and the following results can be generalized to apply to nets

- rather than sequences. If a net of coherent lower previsions converges pointwise to \underline{P} , then \underline{P} is a coherent lower prevision.
4. If we assume only that each \underline{P}_i avoids sure loss then so does \underline{P} , but we need to allow $\underline{P}(X) = -\infty$.
 5. Walley and Fine (1982) develop this frequentist interpretation in the case of finite Ω .

Section 2.7

1. This convention is due to de Finetti. See de Finetti (1974, especially sections 1.9 and 2.5).
2. When \underline{P} is coherent the two definitions of \bar{P} are essentially the same, as any coherent extension of \underline{P} must satisfy $1 - \underline{P}(A^c) = 1 - \underline{P}(1 - A) = -\underline{P}(-A)$.
3. Use 2.4.7(c) with $\mu = 1$.
4. These are the widths of the probability intervals $[\underline{P}(A), \bar{P}(A)]$. (Some authors refer to upper and lower probabilities as ‘interval-valued probabilities’.)
5. Other examples are given in 4.2.3, and by Levi (1974).
6. Alternatively, verify that these upper and lower probabilities satisfy the coherence conditions in Appendix B3.
7. Another way of understanding why the extension is not unique is to note that P_1 and P_2 are *exposed points* with respect to the set of events, meaning that there are events A such that P_j uniquely attains $\underline{P}(A)$, but P_3 and P_4 are not exposed points. In fact, \underline{P} has a unique coherent extension from \mathcal{K} to \mathcal{L} if and only if all the extreme points of $\mathcal{M}(\underline{P})$ are exposed points with respect to \mathcal{K} . (Assuming there are only finitely many extreme points, as for the models in 4.2.)
8. If \underline{P} is defined on a field, coherence can often be established by verifying the stronger property of 2-monotonicity (3.2.4). Compare with the weaker properties in 2.7.4(h).
9. It is possible that $\bar{P}(\bigcup_{j=1}^{\infty} A_j) > \sum_{j=1}^{\infty} \bar{P}(A_j)$. For example, let \bar{P} be the ‘uniform distribution’ on the positive integers (2.9.5), with $A_j = \{j\}$.
10. Here \underline{P} is 3-coherent but not 4-coherent (see Appendix B). Huber (1981, p. 257) gives a similar example.

Section 2.8

1. For the mathematical theory of finitely additive (rather than countably additive) probability, see especially Bhaskara Rao and Bhaskara Rao (1983) and de Finetti (1972, 1974, 1975). Further results are in Dubins and Savage (1965), Dubins (1975), Heath and Sudderth (1978), Sudderth (1980) and Kadane *et al.* (1986).

2. This condition is taken as the definition of linear prevision by de Finetti (1974, section 3.3.5). For equivalent conditions, see de Finetti (1974, section 3.4.1).
3. This is not true without the restriction $\mathcal{K} = -\mathcal{K}$.
4. Alternatively, show that P1 is equivalent to P4 when \mathcal{K} is a linear space and P0 holds, and apply 2.8.3. (Another proof, using 2.8.2, was outlined in 2.3.6.) This is (essentially) the characterization of linear previsions given by de Finetti (1974, section 3.1.5).
5. Axiom P2 is redundant here. It is a good exercise to verify that P2 is a consequence of P0 and P5.
6. These axioms are adopted by Whittle (1970). He adds a continuity axiom that is equivalent to countable additivity of probabilities.
7. When \mathcal{A} is closed under complementation, avoiding sure loss and avoiding sure gain together are equivalent to additivity.
8. If Ω is a ‘large’ space, it may be practically impossible to specify probabilities for all events, and then previsions for other gambles may provide further information. In other problems, previsions may be easier to assess than probabilities.
9. Other versions are $G(-X) = -G(X)$ and $G(A^c) = -G(A)$. See 3.8.3 for equivalent axioms.
10. Keynes (1921, Ch. 3) gives many practical examples where it seems unreasonable to make precise assessments of probabilities.

Section 2.9

1. The uniform distributions on infinite spaces have other defects, discussed in 5.5 and 7.4.
2. This definition can be generalized by replacing P_0 by any coherent lower prevision \underline{P}_0 . This is a simple way of reducing the precision of P_0 , to reflect incomplete confidence in the assessments on which it is based.
3. When X is an event, the tax is a fraction δ of Your reward.
4. Huber and Strassen (1973) and Huber (1965, 1973, 1981) call these gross-error models, and use them to model data generated with probability $1 - \delta$ by a known sampling model, but otherwise generated by a completely unknown mechanism. (They are also known as ε -contamination models.) They have also been used by Bayesian sensitivity analysts as models for imprecise prior beliefs. See Huber (1973, 1981), Berger (1984) and (1985a, sections 3.5.3, 4.7.4), Berger and Berliner (1986), Walley and Pericchi (1988). For a simple example, and comparison with the neighbourhoods defined in the next two examples, see 4.6.5.
5. We ignore the rounding of payouts.
6. The system used at US racetracks is slightly different and can incur sure loss, but only when almost all of the stakes are bet on a single horse.

7. When Ω is infinite, the measures P can be defined in a similar way.
8. This model is discussed by Gerber (1979, Ch. 5). He objects to it because $\bar{P}(A)$ may exceed 1. He also criticizes the extension $\bar{P}(X) = (1 + \delta)P_0(X)$ because it may violate the conditions $\bar{P}(X) \geq P_0(X)$ and $\bar{P}(X + \mu) = \bar{P}(X) + \mu$. The natural extension of the pari-mutuel model, defined in 3.2.5, does satisfy these conditions.
9. See 3.2.5 for an extension of the pari-mutuel upper probability to a coherent upper prevision, defined on all gambles.
10. Compare with the linear–vacuous and pari-mutuel models, for which the differences $\bar{P}(A) - P(A)$ are constant.
11. In the pari-mutuel case, it is more natural to suppose that You act as a bookmaker and sell the gamble A for x . Then the price x is taxed at rate τ . Under the constant odds-ratio model, Your profit $(x - A)^+ = xA^c$ is taxed at rate τ . Again, $xA^c \leq x$.
12. It is a good exercise to prove that the constant odds-ratio lower prevision dominates the linear–vacuous and pari-mutuel (3.2.5) lower previsions for all gambles. An example is given in 4.6.5.
13. More general models, defined in 4.6.4, were studied by DeRobertis and Hartigan (1981).
14. See de Finetti (1972, section 5.17), or Bhaskara Rao and Bhaskara Rao (1983, p. 41).
15. The set of gambles X for which $P_n(X)$ converges is a linear space.
16. Because the uniform distributions P_n are countably additive, this shows that a limit of countably additive distributions need not be countably additive.
17. De Finetti (1972, section 5.17) does regard these dominating linear previsions as reasonable models for choosing a positive integer ‘at random’. He seems to regard other models as reasonable, provided they satisfy $P(A) = 0$ for all finite A .
18. Billingsley (1979, p. 37). There are other finitely additive probabilities on B that are translation-invariant (see note 11 of section 3.5).
19. There is a simple proof of this, given in Billingsley (1979) and Royden (1963), which uses the axiom of choice to define a countable partition of Ω , all of whose sets are translates of each other. Assuming the continuum hypothesis as well as the axiom of choice, Ulam (1930) proved that there is no countably additive probability on all subsets of Ω that assigns zero probability to every singleton. It follows that Lebesgue measure has no countably additive extension to all subsets, even without requiring translation-invariance. The existence of non-measurable sets can be proved from assumptions that are considerably weaker than the axiom of choice (see 3.6.9).
20. This agrees with the more usual definition of outer measure, due to

- Carathéodory, because the outer measure is regular. In the formulas for outer or inner measure, B can be restricted to be open or closed respectively.
21. This is equivalent to the more usual definition of measurable sets, as those A such that $\bar{P}(B) = \bar{P}(A \cap B) + \bar{P}(A^c \cap B)$ for all subsets B . The class of Lebesgue-measurable sets is the smallest σ -field containing the Borel sets and the sets of outer measure zero (property 5 of 3.2.7).
 22. Such a set A is defined in 3.6.9.
 23. All the linear previsions that dominate \underline{P} or \bar{Q}' are translation-invariant, but (unlike 2.9.5) there are translation-invariant linear previsions that do not dominate \bar{Q}' .
 24. Alternatively, verify that the natural extension of \underline{P} satisfies axioms P1–P3.
 25. See 3.6.8. When Ω is finite, define A_0 to be the intersection of all sets in \mathcal{A} . The 0–1 valued additive probabilities are degenerate at a single point when Ω is finite (A_0 is a singleton set).
 26. Another example is discussed by de Finetti (1972, p. 94).

Section 2.10

1. Giles (1976) has presented a thoroughly operational theory of subjective probability, in which probability judgements, made during a ‘dialogue’ between You and an opponent, are identified with commitments to accept specific gambles from the opponent. Two operational definitions of prevision are suggested by de Finetti (1974, section 3.3), see also Lad (1990).
2. Many incomplete models \underline{P} can correctly describe the same beliefs. (If \underline{P} does, so does any $\underline{Q} \leq \underline{P}$.)
3. This seems to be the interpretation of Williams (1976) and Levi (1980). (Williams, in his note 6, recognizes that $\underline{P}(X)$ is not exactly determined.)
4. Beliefs that are correctly described by \underline{P} can also be correctly described by any less precise model \underline{Q} (provided this is not regarded as an exhaustive model). Bayesians often approximate an additive probability P by a more tractable one Q , but this kind of approximation does not preserve correctness.
5. Precise probability models are always exhaustive.
6. This type of indeterminacy is discussed by Russell (1948, p. 276).
7. See Fishburn (1964, 1965), Suppes (1974), Dickey (1980) and Mellor (1980). Lindley, Tversky and Brown (1979) construct a theoretical model for errors in elicitation, although it is not clear whether they intend the underlying P_T to be interpreted descriptively or normatively. Various kinds of incomplete elicitation are described in 5.2.

8. Mellor (1980) discusses the relation between conscious and unconscious dispositions.
9. This is suggested by Jeffreys (1931, section 2.5) and Good (1962a, 1965, 1980). Giles (1976, pp. 67–8) suggests a version of the sensitivity analysis interpretation in which P_T represents unknown aleatory probabilities.
10. This seems to be the interpretation of Berger (1984, pp. 71–3), Berger and Wolpert (1984, p. 136), and perhaps Lindley, Tversky and Brown (1979).
11. For further discussion of sensitivity analysis interpretations, see 5.9.
12. For examples and discussion of the difference, see Walley and Pericchi (1988).

Section 2.11

1. Berger (1985a, sections 3.7 and 4.12) discusses some other objections.
2. For examples of how personal judgements can be supported through detailed description of the evidence, see the case studies of Mosteller and Wallace (1964) or Walley and Campello de Souza (1986). See section 1.7 for further discussion.
3. ‘Objective’ is sometimes used in yet another sense, e.g. by Berger (1985a), to describe statistical procedures which can be used automatically, without any prior assessments or other judgements. Bayesian noninformative priors are objective in this sense, but apparently not in any of the four senses discussed here (see 5.5).
4. Such careful authors as Savage (1972a) and de Finetti (1974, 1975) confuse physicality with both interpersonal agreement and certainty about physical reality. They argue that we can never justify certainty about the values of theoretical quantities, whose ‘objectivity’ is therefore either illusory or reducible to interpersonal agreement! For a more illuminating discussion of interpersonal agreement, see Dawid (1982a). (Compare with the ‘intersubjective values’ discussed by Darwall (1983).) Other discussions of objectivity from a Bayesian point of view include Jeffreys (1931), Box and Tiao (1973), Good (1976) and Berger (1985a, section 3.7).
5. Agassi (1975) distinguishes physicality (which he calls ‘the real’) from interpersonal agreement (‘the common’). See also Popper (1959a) and Mellor (1971).
6. This is not to deny that there are serious difficulties in developing an aleatory interpretation (see 7.2).
7. This is the view of de Finetti (1974, 1975) and Dawid (1982a). It agrees with Peirce’s view that ‘the objectivity of truth really consists in the fact that, in the end, every sincere inquirer will be led to embrace it’ (1878, p. 288). Rorty (1980, p. 333) also identifies ‘truth’ with ‘interpersonal agreement’. Compare with the account of truth as ‘correspondence with reality’ that is implicit in the first sense of objectivity.
8. Cox and Hinkley (1974, p. 389) judge the personalist Bayesian approach

- to be ‘inapplicable because it treats information derived from data as on exactly equal footing with probabilities derived from vague and unspecified sources’. Fine (1973, p. 231) makes a similar criticism.
9. As suggested by Good (1950, section 7.4).
 10. They can make use of non-statistical information in an informal way, e.g. in choosing critical levels and in interpreting numerical results, but the frequentist theory gives no guidance on how to do so. See Hodges and Lehmann (1952).
 11. See, for example, Barnard’s discussion of Smith (1961).
 12. This objection is raised by Fraser (1977) and Berger (1985a, p. 260). It is a valid objection to de Finetti’s (1974) account, on which You must offer two-sided betting rates and Your opponent can choose which side of the bet to take (see Appendix H2).
 13. Kyburg (1978) argues that people will naturally ‘avoid sure loss’ in the practical sense, even when they ‘incur sure loss’ in the technical sense, by refusing to combine gambles in this way.
 14. There is a close connection between coherence and the decision-theoretic criterion of admissibility. This is explained in 3.9, and in de Finetti (1974, section 3.3).
 15. More realistic experiments are described by Winkler (1971), Beach (1975), Hogarth (1975, 1980), Slovic *et al.* (1977), Hogarth and Kunreuther (1985), and von Winterfeldt and Edwards (1986).
 16. See, for example, Tversky (1969), Edwards (1968), Cohen (1981), Kahneman *et al.* (1982), Slovic and Lichtenstein (1983), as well as the preceding references.
 17. Fine (1988, p. 399) suggests that several probabilistic judgements may each be soundly based on evidence, but the judgements together may incur sure loss. In his words, ‘our best knowledge may be technically incoherent’. He argues that our current state of knowledge concerning the physical process of flicker noise supports judgements that the process is stationary and has bounded but divergent time averages, which cannot be modelled by a coherent probability model.
 18. De Finetti (1975, Appendices 2–13) discusses these issues at length. He admits only observable spaces.
 19. The existence of such a verification procedure was proposed by the logical positivists as a general criterion for testing whether a statement is meaningful. See the papers in Ayer (1959), Carnap (1936–37), Ayer (1973) and Popper (1962, Ch. 11). Later versions of the verification principle, e.g. Carnap (1956) and Braithwaite (1953, Ch. 6), did admit hypotheses about unobservable states as meaningful, provided there are possible observations that can ‘confirm’ or ‘support’ them to some degree.
 20. The implications of beliefs about Θ for beliefs about statistical observables can be calculated by the type of natural extension studied in 6.7.

Chapter 3

Section 3.1

1. This was recognized by Williams (1976, p. 239).
2. Formulas (b) and (e) were stated by Huber (1981, p. 258) and Williams (1976) respectively, for the special case where \mathcal{K} contains only events.
3. See Halmos (1950) or other texts on measure theory. The example of Lebesgue measure is examined in 3.2.7.
4. If $\mathcal{K} = -\mathcal{K}$ and \underline{P} is linear then $\underline{E}(Z) = \bar{E}(Z)$, by 3.1.3(c). In general, $\underline{E}(Z) \leq \underline{E}(Z) \leq \bar{E}(Z)$.
5. When \mathcal{K} is finite we can set $\delta = 0$.

Section 3.2

1. For studies of 2-monotonicity, see Choquet (1953–54), Huber (1973), Huber and Strassen (1973), Anger (1977), Huber (1981, Ch. 10) and Walley (1981, Ch. 6).
2. See Huber (1981, pp. 260–2) or Walley (1981) for a proof. These formulas hold for all X in $\mathcal{K}(\mathcal{A})$ if and only if \underline{P} is 2-monotone. (More generally, the expressions give a lower bound for $\underline{E}(X)$. If 2-monotonicity fails for events A and B then there is strict inequality when $X = A + B$.) The natural extension of \underline{P} to all gambles is determined by these formulas and 3.1.4. Alternatively, use 3.1.5 to extend \underline{P} to all subsets of Ω (this extension preserves 2-monotonicity), and apply the formulas to obtain \underline{E} on \mathcal{L} . If \underline{P} is 2-monotone, \underline{E} has the generalized 2-monotonicity property $\underline{E}(X \vee Y) + \underline{E}(X \wedge Y) \geq \underline{E}(X) + \underline{E}(Y)$, where \vee and \wedge denote maximum and minimum. When \mathcal{A} is finite and \underline{P} is completely monotone on \mathcal{A} (defined in 5.13.1), its natural extension is defined by $\underline{E}(X) = \sum_{A \in \mathcal{A}} m(A) \inf\{X(\omega) : \omega \in A\}$, where m is the probability assignment for \underline{P} .
3. In general, $\bar{P}(X) = (1-\varepsilon)P_0(X|X > x_t) + \varepsilon x_t$, where $\varepsilon = 1 - (1+\delta)P_0(X > x_t)$.
4. It is a good exercise to derive these results directly from the definition of \underline{E} (3.1.1) and the properties of filters, without using 2-monotonicity. For this example, $\underline{E}(X) = \underline{E}(X)$ in 3.1.6, so that \underline{E} is the unique coherent extension of \underline{P} to \mathcal{L} .

Section 3.3

1. When both Ω and \mathcal{K} are finite (as for the finitely generated models in 4.2), the separating hyperplane theorem can be replaced by a finite-dimensional version such as the theorem of the alternative (Appendix E). See Theorem 2.1 of Walley (1981).
2. This is essentially contained in Heath and Sudderth (1978). The gambles

- in \mathcal{D} and \mathcal{V} can be interpreted as almost desirable, and then (a) asserts that \mathcal{D} avoids sure loss in the sense of 3.7.1.
3. Versions of this result for finite Ω were given by Smith (1961, p. 11) and Huber (1981, p. 259). More general versions are in Williams (1976), Giles (1976) and Walley (1981). The linear previsions P in $\mathcal{M}(P)$ are called *medial* probabilities by Smith, because they satisfy $\underline{P} \leq P \leq \bar{P}$.
 4. It is clear also from the equivalence of (b) and (c) in 3.3.2 that the domain of the dominating linear previsions is immaterial.

Section 3.4

1. This result shows that the natural extension is ‘natural’ also under a sensitivity analysis interpretation. (Since $\mathcal{M}(\underline{E}) = \mathcal{M}(\underline{P})$, \underline{E} provides exactly the same information as \underline{P} about the ideal linear prevision.) For similar results, see Williams (1976, p. 240) and Huber (1981, p. 258).
2. Theorem 3.3.3(a), which relies on the separating hyperplane theorem, is needed to prove that linear extensions exist. Compare with the elementary result 3.1.2, which shows that all coherent lower previsions (including linear previsions) can be extended to \mathcal{L} as coherent lower previsions.
3. Generally $\mathcal{M}(P) \supset \mathcal{M}(P')$, but not all Q in $\mathcal{M}(P)$ are extensions of P . For example, let A be a non-trivial event, $\mathcal{K} = \{A\}$ and $P(A) = 0$. Then $\mathcal{M}(P)$ contains all linear previsions but $\mathcal{M}(P')$ contains only those Q with $Q(A) = 0$.
4. When P is an additive probability defined on \mathcal{A} , the restriction $\mathcal{K} = -\mathcal{K}$ in 3.4.1 and 3.4.2 can be replaced by $\mathcal{A} = \mathcal{A}^c$.
5. See also de Finetti (1937, p. 108). We regard Theorems 3.1.2 and 3.4.1, which characterize the natural extension, as more fundamental than 3.4.2 and 3.4.3, which characterize the class of linear extensions. The problem of calculating the implications of given (precise) probabilities for the probabilities of other events was considered much earlier, by Boole (1854), who also regarded it as the central problem of probability theory. He proposed a general, but complicated, method for solving the problem. The method, which appears to be closely related to natural extension, is explained by Hailperin (1976).
6. This is clear also from 3.1.8, as $\underline{E}(Z) = \bar{E}(Z)$. See also de Finetti (1972, sections 5.10 and 6.5) and (1975, Appendix 15).
7. It is not sufficient here that \underline{P} avoids sure loss.

Section 3.5

1. This formulation is more general than it appears, as any semigroup \mathcal{G} can be regarded as a semigroup of transformations of itself. (Regard $g \in \mathcal{G}$ as a mapping from \mathcal{G} into \mathcal{G} defined by $g(h) = g * h$, where $*$ is the semigroup

operation. Composition is defined by $g_1 g_2 = g_1 * g_2$.) So we can apply the following results to any semigroup \mathcal{G} by taking $\Omega = \mathcal{G}$.

2. Natural extension preserves \mathcal{G} -invariance of \underline{P} whenever \mathcal{G} is a group (it suffices that all elements of \mathcal{G} have right-inverses), but not necessarily when \mathcal{G} is a semigroup. (Define \underline{P} only on the constant gambles in 3.5.7. Its natural extension is the vacuous lower prevision, but this is not \mathcal{G} -invariant.)
3. A semigroup \mathcal{G} is called (left-) *amenable* when there is a \mathcal{G} -invariant linear prevision on $\mathcal{L}(\mathcal{G})$. Condition (a) yields a necessary and sufficient condition for amenability, by taking $\Omega = \mathcal{G}$. See Greenleaf (1969, p. 4).

4. For this result, see Day (1942) or Royden (1963, Ch. 10).

5. By taking $\Omega = \mathcal{G}$, this proves that every Abelian semigroup is amenable (Day, 1942). See Granirer (1973) for a similar result in the case where $\Omega = \mathcal{G}$ is an amenable, locally compact topological group (not necessarily Abelian).

6. Here \mathcal{G} is not a group. (Its elements have left-inverses but not right-inverses.)

7. See Berberian (1974, p. 121) or Bhaskara Rao and Bhaskara Rao (1983, p. 39).

8. The Banach limit of a sequence X is unique just when $\underline{Q}(X) = \bar{Q}(X)$.

9. Note that $\underline{Q}(A) = 0$ just when \mathbb{R} cannot be covered by finitely many translates of A . We can construct A to be a countable union of intervals, with this property but with $\underline{Q}'(A) = 1$. There are also sets A (countable unions of intervals) with $\underline{P}(A) = 1$ but $\underline{Q}'(A) = 0$.

10. Similarly, since the translation group on \mathbb{R}^n is Abelian, there are translation-invariant linear previsions on \mathbb{R}^n . Invariance under the larger (non-Abelian) group of all isometries is a different matter. (Isometries are transformations which preserve distances between points, such as translations, reflections or rotations.) There are isometry-invariant linear previsions on \mathbb{R}^n if $n = 1$ or $n = 2$, but not if $n \geq 3$. See Moore (1982, sections 3.7 and 4.11).

11. There are translation-invariant linear previsions on $\mathcal{L}(\Omega)$ that are not extensions of Lebesgue measure. By 3.5.6, their lower envelope is $\underline{Q}^*(X) = \sup \{\inf n^{-1} \sum_{j=1}^n X \oplus c_j : n \geq 1, 0 \leq c_j < 1\} \leq \underline{Q}(X)$. We can construct sets A (countable unions of intervals) such that $\underline{Q}^*(A) = 1$ but $v(A)$ is arbitrarily small. This means that there are translation-invariant linear previsions on $\mathcal{L}(\Omega)$ which assign probability 1 to Borel sets whose Lebesgue measure is arbitrarily small. (So Lebesgue measure is not the only finitely additive, translation-invariant probability on the Borel sets.)

Section 3.6

1. Whereas sections 3.3–3.5 used only the separating hyperplane theorem, in this section we need stronger results from topology: Tychonoff's compactness theorem (which is used in Appendix D4 to show that \mathcal{P} is weak*-compact), the strong separation theorem (Appendix E3) and the

Krein–Milman theorem (E5). See note 12 concerning the strength of these theorems.

2. $\mathcal{M}(\underline{P})$ is closed, but not necessarily compact, under two other natural topologies on \mathcal{P} , the norm (or strong or total variation) topology and the weak topology, which are each stronger (i.e. have more closed sets) than the weak* topology. It is the compactness of \mathcal{P} , hence of $\mathcal{M}(\underline{P})$, that makes the weak* topology especially useful.

The properties of compactness and convexity of \mathcal{M} do not appear to have any behavioural significance. All classes \mathcal{M} which have the same lower envelope \underline{P} convey the same information about behavioural dispositions. So we regard $\mathcal{M}(\underline{P})$, its set of extreme points (which is not convex), and its relative interior (which is not closed) as equivalent models. A sensitivity analyst might want to distinguish between these models.

3. Huber (1981, section 10.2) proves essentially this result for the case of finite Ω .

4. Allowing \underline{P} to be defined on a smaller domain, or to be incoherent, does not introduce any new classes $\mathcal{M}(\underline{P})$, because either $\mathcal{M}(\underline{P})$ is empty (if \underline{P} incurs sure loss) or $\mathcal{M}(\underline{P}) = \mathcal{M}(\underline{E})$ where the natural extension \underline{E} is coherent on \mathcal{L} (3.4.1).

5. If Ω is finite or \underline{P} is finitely generated (4.2) then $\mathcal{M}(\underline{P})$ is the convex hull of $\text{ext } \mathcal{M}(\underline{P})$. (The operation of closure is not needed.)

6. To prove this, apply (c) of the Krein–Milman theorem to the evaluation functional X^* defined by $X^*(P) = P(X)$, which is a weak*-continuous linear functional (Appendix D3). The closure of $\text{ext } \mathcal{M}(\underline{P})$ is the smallest weak*-compact set with lower envelope \underline{P} . See Day (1973, p. 103) or Holmes (1975, p. 74).

7. Another characterization of \mathcal{Q} is that it is the class of all linear previsions P such that $P(XY) = P(X)P(Y)$ for all gambles X and Y .

8. This and the following results are trivial when Ω is finite, since then all filters have the form $\mathcal{A} = \{A : A \supset A_0\}$ for some event A_0 , and all ultrafilters have the form $\mathcal{A} = \{A : \omega_0 \in A\}$ for some $\omega_0 \in \Omega$.

9. See Choquet (1953–54, p. 245).

10. They are also called principal ultrafilters.

11. In general, \mathcal{Q} is the weak*-closure of the set of degenerate linear previsions.

12. The standard example of a non-measurable set invokes the axiom of choice. The results of Solovay (1970) imply that the existence of non-measurable sets cannot be deduced from the axioms of Zermelo–Fraenkel set theory without further assumptions. But it is sufficient to assume the special case of the weak ultrafilter theorem with $\Omega = \mathbb{Z}^+$, which is considerably weaker than the axiom of choice. Moore (1982) lists other sufficient conditions. (All these conditions are non-constructive: no non-measurable set has been explicitly specified.)

We can summarize the implications amongst the theorems used in this chapter as follows. Write AC for the axiom of choice; KM for the Krein–Milman theorem (Appendix E5(a)); UT for the ultrafilter theorem (3.6.5), which is equivalent to various other well-known results such as Tychonoff's compactness theorem for Hausdorff spaces, the Boolean prime ideal theorem and Stone's representation theorem for Boolean algebras; SH for the separating hyperplane theorem (Appendix E1), which is equivalent to the Hahn–Banach theorem; WUT for the weak ultrafilter theorem (3.6.8); and NM for the existence of subsets of $[0, 1]$ that are not Lebesgue measurable (3.6.9). Assuming Zermelo–Fraenkel set theory, we use \Leftrightarrow to mean ‘is equivalent to’ and \Rightarrow to mean ‘is strictly stronger than’ (i.e. implies, but is not implied by).

Then $AC \Leftrightarrow (KM + UT) \Rightarrow UT \Rightarrow SH$ and $UT \Rightarrow WUT \Rightarrow NM$.

So AC is considerably stronger than both SH and NM. For further details and related results, see Jech (1973), Pincus (1974) and Moore (1982).

13. This proof is due to Sierpinski (1938). The set A is ‘extremely non-measurable’ in the sense that $\bar{P}(A \cap S) = v(S)$ and $\bar{P}(A \cap S) = 0$ for every Lebesgue-measurable set S . See also Halmos (1950, p. 70).
14. Here A and C partition B^c so that (i) any x and y such that $x + y \in B$ lie in different sets, and (ii) any x and y such that $x - y \in B$ lie in the same set.
15. That can be proved using finite sub-additivity and translation-invariance of the outer measure \bar{P} .

Section 3.7

1. This is just the condition in the separation lemma 3.3.2.
2. For finitely generated models (with Ω and \mathcal{D} both finite), the operation of closure is not needed in defining \mathcal{E} .
3. Avoiding sure loss can also be characterized in terms of \mathcal{E} : \mathcal{D} avoids sure loss if and only if \mathcal{E} is a proper subset of \mathcal{L} . (Equivalently, \mathcal{E} does not contain the gamble -1 .)
4. The first four axioms are essentially the same as those suggested by Williams (1976, p. 241). Coherence can be characterized under the weaker assumption that \mathcal{K} is a convex cone, by replacing D1 by monotonicity and non-triviality axioms.
5. If \mathcal{D} and \underline{P} correspond as in 3.8.1 then so do their natural extensions \mathcal{E} and \underline{E} . So the two definitions of natural extension in 3.1.1 and 3.7.2 are essentially the same. Coherence is preserved under arbitrary intersections of coherent classes \mathcal{D} , which corresponds to forming lower envelopes of coherent lower previsions.
6. This property is called preference-reversal coherence by Buehler (1976).
7. These are slightly different from (but apparently equivalent to) axioms suggested by Giles (1976) and by Giron and Rios (1980).

8. For the theory of partially ordered linear spaces, see Jameson (1970), Wong and Ng (1973), or Kelley and Namioka (1963).
9. The first four axioms are essentially those required by Williams (1975b), except that he assumes the zero gamble is in \mathcal{D}^+ . Smith (1961, p. 15) requires \mathcal{D}^+ to be an open convex cone containing all positive gambles.
10. It is tempting to define strict preference in terms of almost-preference, by $X > Y$ when $X \geq Y$ and not $Y \geq X$. However, this does not guarantee (under our interpretation) that X is preferred to Y . Similarly, we could define an equivalence relation of almost-indifference by $X \approx Y$ when $X \geq Y$ and $Y \geq X$, but this cannot be interpreted as indifference between X and Y . For example, if X is the surely negative gamble in Example 2.4.5 then $X \approx 0$, but You would prefer 0 to X .

Section 3.8

1. The correspondences hold more generally, for coherent lower previsions on an arbitrary domain \mathcal{K} , provided the domains in 2–5 are defined appropriately.
2. Some of the other correspondences were defined in 3.7.
3. $\mathcal{M}(\mathcal{D})$ is often called the polar of \mathcal{D} .
4. Of course the models have been defined to achieve this. More detailed or informative models are possible if the constraints we imposed are relaxed. In particular, a class of really desirable gambles need not satisfy the closure axiom D4 or the openness axiom D7 (see Appendix F). Under a sensitivity analysis interpretation, \mathcal{M} need not be convex or compact. Sensitivity analysts might want to distinguish different classes \mathcal{M} which have the same lower envelope. For example, the set of all degenerate linear previsions seems to be a very different model from its complementary set, but both have the vacuous lower revision as their lower envelope.
5. See 4.2.3, and other examples in Chapter 4.
6. But the coherence axioms for conditional previsions may be simplest in terms of real desirability (Appendix F).
7. An alternative is to allow $\underline{P}(X)$ to take values in the nonstandard real numbers. See de Finetti (1972, section 5.13), Parikh and Parnes (1974), or Davis (1977).
8. An alternative approach, which we follow in Chapters 6 and 7, is to elicit the extra information directly through assessments of conditional previsions. It seems clear that uncertainty cannot always be adequately described through unconditional previsions alone.

Section 3.9

1. Concerning decision making with imprecise probabilities and utilities, see Smith (1961), Fishburn (1964, 1965, 1975), Levi (1974, 1980, 1982), Giron and Rios (1980), Kmietowicz and Pearman (1981), Wolfson and Fine

(1982), Gardenfors and Sahlin (1982, 1983), Berger (1984, 1985a), and other references in Berger (1984).

2. For a practical application in which the model of precise utilities but imprecise probabilities is realistic, see Walley and Campello de Souza (1986), concerning a decision about whether to buy a solar water-heating system.
3. It may be useful to consider several possibility spaces and combine the resulting preferences.

4. There may be several actions which correspond to the same gamble, and You may have preferences between these. To express these preferences, You must enlarge the space Ω so that the two actions correspond to different gambles. An example is given in note 13.

5. Use the finite intersection property to show that any subset of \mathcal{K} that is completely ordered by \geq has an upper bound in \mathcal{K} . By Zorn's lemma, there is some Y in \mathcal{K} that is maximal under \geq , and Y is admissible. Next suppose that P is coherent, hence sup-norm continuous. So the set \mathcal{K}' containing all X in \mathcal{K} which maximize $P(X)$ is a non-empty compact subset of \mathcal{K} . Apply the first result to \mathcal{K}' to conclude that there are admissible minimax gambles in \mathcal{K} . By the argument in 3.9.7, these gambles are maximal.

6. These issues are discussed in 5.6.

7. If Ω is finite and each element has positive probability then every Bayes gamble under P is maximal. In general, some of the Bayes gambles may be inadmissible.

8. If \mathcal{K}_0 is compact (e.g. finite), so is its convex hull \mathcal{K} , and there are maximal gambles in \mathcal{K} .

9. Assuming \mathcal{K} is compact (in the supremum-norm topology) as well as convex, an alternative proof is illuminating. Defining $\Pi(Y, P) = P(X - Y)$, $\mathcal{V} = \mathcal{K}$ and $\mathcal{W} = \mathcal{M}(P)$, verify that the conditions of the minimax theorem (Appendix E6) are satisfied. So

$$\min_{Y \in \mathcal{K}} \max_{P \in \mathcal{M}(P)} P(X - Y) = \min_{Y \in \mathcal{K}} \bar{P}(X - Y) = \max_{P \in \mathcal{M}(P)} \min_{Y \in \mathcal{K}} P(X - Y).$$

The last expression is non-negative if and only if there is $P \in \mathcal{M}(P)$ under which X is a Bayes gamble. That establishes the theorem.

10. The term 'maximin' is perhaps more appropriate, because X_0 maximizes the minimum expected utility. The term 'minimax' comes from statistical decision theory, where loss (negative utility) is used: X_0 minimizes the maximum expected loss. Alternatively, one might minimize the maximum expected regret. If $Z(\omega) = \max\{X(\omega): X \in \mathcal{K}\}$ is the best possible reward when ω is the true state, the regret from choosing X is $Z - X$. This leads to the minimax regret rule: choose an admissible X from \mathcal{K} to maximize $P(X - Z)$. The minimax loss and minimax regret rules can produce different decisions.

11. Some of the defects of minimax rules are discussed in Berger (1985a, Ch. 5).
12. This is well known in the case of finite Ω (see Ferguson, 1967, or Berger, 1985a), and it holds generally for bounded loss functions. (Of course, not all admissible rules are Bayes decision rules under some countably additive prior prevision.)

This kind of result is often invoked to justify the Bayesian approach, for both decision making and statistical inference. Berger (1985a, section 4.8.3) even suggests that admissibility is a more fundamental criterion than coherence. But admissibility is essentially a decision-theoretic criterion: it refers to a specific set of utility functions. It is relevant in a specific decision problem, but it has no obvious relevance to problems of inference. Coherence, which applies to both inference and decision, therefore seems to be more fundamental and more general than admissibility. We regard admissibility as a consequence of coherence that applies to the special type of decision problem in which actions are evaluated through precise utility functions.

13. See especially Ferguson (1967). In classical decision theory, Ω is taken to be the parameter space Θ , but a larger space may be needed in order to distinguish between decision rules with the same risk function. Suppose, for example, that d_1 and d_2 are identical except when x is observed, and then d_1 results in greater utility than d_2 whatever the value of θ . Then d_1 should be preferred to d_2 . But if x has aleatory probability zero for each θ then d_1 and d_2 have identical risk functions, and a classical decision theorist cannot distinguish between them. They can be distinguished when Ω is extended to the product space $\Theta \times \mathcal{X}$ (d_2 is inadmissible when evaluated in the product space). In this example, d_1 is at least as good as d_2 conditional on each possible observation, and d_1 is strictly better than d_2 conditional on x . A general criterion is: if X is at least as good as Y conditional on B , for every B in a partition \mathcal{B} of Ω , and X is strictly better than Y conditional on some B , then X should be preferred to Y . This yields the admissibility criterion as a special case when \mathcal{B} corresponds to Θ . (As suggested by the example, You could also let \mathcal{B} correspond to \mathcal{K} and evaluate preferences conditional on the observation.)

14. In statistical decision problems, the admissible rules in the class of all non-randomized rules need not be Bayes decision rules.
15. The imprecise utilities could be assessed through judgements of preferences between *lotteries*, in which You receive known consequences with known chances. The given partial ordering can then be constructed directly by natural extension of these judgements, without reference to the precise probability ρ and utility μ .

Chapter 4

Section 4.1

1. Good (1962a) has suggested a similar procedure, which is outlined in 5.9.2. Good translates judgement 2 of 4.1.1 into the constraint $\underline{P}(A) \geq \bar{P}(B)$, which is stronger than our translation $\underline{P}(A - B) \geq 0$. Similarly, Good's translation of judgement 6 is different from ours.
2. Other examples, including imprecise assessments of density functions and distribution functions, are given in 4.6.
3. For example, a judgement that $\underline{P}(X) \leq 2$ is not a direct judgement of desirability. (Nor can it be regarded, by sensitivity analysts, as a constraint on the true linear prevision.)
4. The simplest way to model a judgement that X is really desirable is to include X in \mathcal{D} , but this loses some information. Another way is to find some positive δ such that $X - \delta$ is almost desirable.
5. See the models in 5.3 and 5.4 for relationships between these quantities.
6. A judgement that the lower variance of X is at least 3 means that the gamble $(X - \mu)^2 - 3$ is almost desirable for every real μ . This fits into the general elicitation procedure, but a judgement that the upper variance is no greater than 3 does not: it means that a gamble $3 - (X - \mu)^2$ is almost desirable for some unspecified value of μ . This difference arises because the variance functional Υ , defined by $\Upsilon(P) = P(X^2) - (P(X))^2$, is concave; see Appendix G. The class of linear previsions P satisfying $\Upsilon(P) \geq 3$ is convex, but the class satisfying $\Upsilon(P) \leq 3$ is not. Sensitivity analysts can admit both kinds of judgements, by regarding them as constraints on the true P . By specifying upper bounds for variances, they can obtain models \mathcal{M} that are not convex.
7. For example, the constraint that all distributions in \mathcal{M} are symmetric about a fixed point will usually be unreasonable, because it is equivalent to many precise assessments of previsions; see Pericchi and Walley (1989).
8. The individual judgements of desirability are combined here by the conjunction rule (4.3.8), which forms the intersection of the half-spaces \mathcal{M}_x that represent each judgement.
9. The lower prevision \underline{P} is not determined by these upper and lower probabilities; see 4.2.3 (and Example 2.7.3).
10. If You have been honest, these models represent Your judgements about Your own beliefs. Mellor (1980) calls them Your 'degrees of assent'. See also de Sousa (1971). It is possible, of course, for You to be mistaken about Your beliefs, so the elicited model can be incorrect. Any measurement technique is fallible. We expect the general elicitation procedure to be more reliable than Bayesian elicitation, however, because You can abstain from making judgements whenever You are unsure about Your beliefs and preferences.

Some more elaborate, second-order probability models for 'beliefs about beliefs' are described in 5.10.

11. See especially de Finetti (1974, Ch. 3) and Giles (1976).
12. Although there is nothing in our model to prevent this.
13. The literature on Bayesian elicitation includes Berger (1985a, Ch. 3), de Finetti (1972, 1974), Hampton, Moore and Thomas (1973), Hogarth (1975, 1980), Huber (1974), Spetzler and Stael von Holstein (1975), Slovic, Fischhoff and Lichtenstein (1977), von Winterfeldt and Edwards (1986, Ch. 4), Wallsten and Budescu (1983), and Winkler (1967a, 1967b). The choice of domain \mathcal{K} , on which precise previsions are assessed, is important in Bayesian elicitation. In de Finetti's approach, linear previsions must be assessed for all gambles in \mathcal{K} , but no additional assessments (however imprecise) can be made concerning the other gambles. The Bayesian analysis, which assumes linearity, is confined to gambles in the linear space generated by \mathcal{K} . (Other Bayesian approaches are more restrictive in requiring \mathcal{K} to be a linear space or a field of sets.) Our analysis is not restricted to any special class of gambles, since imprecise previsions are determined for all gambles by natural extension.
14. See Winkler (1967a) for evidence of this.

Section 4.2

1. It is quite possible to make infinitely many judgements and to evaluate gambles that are not simple. For example, You might judge that the singleton $\{n\}$ is at least as probable as $\{n + 1\}$ for every positive integer n . When Ω is the unit interval, You may be able to assess upper and lower previsions for $X(\omega) = \omega$, which is not a simple gamble.
2. See Gale (1960, Theorem 2.13), or Rockafellar (1970).
3. Equivalently, the dual system $\sum_{j=1}^n \lambda_j X_j = -1$ and $\lambda_j \geq 0$ for $1 \leq j \leq n$ has no solution $(\lambda_1, \dots, \lambda_n)$. This dual problem is discussed in Appendix A.
4. One way to compute the extreme points of \mathcal{M} is to consider all subsets J of $\{1, 2, \dots, n\}$ which have size $k - 1$ and are such that $\{X_j : j \in J\}$ and the constant gamble 1 are linearly independent. For each such J , compute the unique solution P of the k linear equations $P(X_j) = 0$ for each $j \in J$ and $P(1) = 1$. The extreme points are those P which satisfy all n inequalities $P(X_i) \geq 0$. (In the simplex representation, each extreme point is the point at which $k - 1$ hyperplanes intersect.) See the following example (4.2.2) for illustration. A new algorithm for solving linear programming problems in polynomial time is described by Karmarkar (1984). This may be useful for computing \underline{P} when Ω and \mathcal{D} are large sets.
5. This is described by Dempster (1967a) and de Finetti (1972). There are two other geometric representations that can be useful. The first one, like the simplex representation, identifies linear previsions with points, and

gambles with hyperplanes. Suppose that $\mathcal{D} = \{X_1, X_2, \dots, X_m\}$. The possible states $\omega \in \Omega$ correspond to the possible reward-vectors $(X_1(\omega), \dots, X_m(\omega))$, which can be plotted as points in \mathbb{R}^m . The linear previsions $P \in \mathcal{P}$ correspond to points $(P(X_1), \dots, P(X_m))$ in \mathcal{V} , the closed convex hull of the set of possible reward-vectors. Then $\mathcal{M}(\mathcal{D})$ is just the intersection of \mathcal{V} with the non-negative orthant, and \mathcal{D} avoids sure loss just when this intersection is non-empty, i.e. when some convex combination of possible reward-vectors has all its coordinates non-negative (see 3.3.2). Otherwise, there is a hyperplane $\sum_{j=1}^m \lambda_j x_j = \mu$ which separates the possible reward-vectors from the non-negative orthant, so that $\lambda_j \geq 0$ and $\mu < 0$, and this represents a gamble $\sum_{j=1}^m \lambda_j X_j$ which is a sure loss. This representation is discussed by de Finetti (1974, Ch. 3). It is especially useful when \mathcal{D} is small, and for showing the effect of enlarging or reducing Ω (4.3.6, 4.3.7).

The dual of the simplex representation identifies gambles with points and linear previsions with hyperplanes. When $\Omega = \{\omega_1, \dots, \omega_k\}$, represent each gamble X by the point $(X(\omega_1), \dots, X(\omega_k))$ in \mathbb{R}^k . Then \mathcal{D} is a finite set of points and its natural extension \mathcal{E} is the smallest convex cone containing \mathcal{D} and all points in the non-negative orthant. Linear previsions correspond to hyperplanes separating the non-negative orthant from the non-positive orthant, and \mathcal{M} consists of all such hyperplanes that also separate \mathcal{D} from the non-positive orthant. This k -dimensional representation can be reduced to $k - 1$ dimensions by identifying all gambles λX for which λ is positive. (If any one is desirable, so are the others.) This enables a three-point space Ω to be represented in two dimensions, as in the simplex representation. An example is given by Walley (1984–85). This representation is especially useful for constructing the class \mathcal{E} of desirable gambles.

Section 4.3

1. The assessment process could be greatly facilitated by using a suitable *expert system*, a computer program which accepts various types of judgements from its user, computes their natural extension, and displays summaries of the current model after each judgement. For example, the system might display the current model graphically, through simplex or other representations, or give summaries of its precision, such as the size of \mathcal{M} or degrees of imprecision $\Delta(X)$ for important variables, or, in a decision problem, compute the current set of maximal actions. You would use these summaries to decide whether previous assessments are reasonable and what kind of assessments should be considered next. Ideally, the expert system would guide this process by asking You to consider particular types of gambles and assessments, in order to achieve goals such as reducing the set of maximal actions. It would be quite easy to design a system that could elicit probabilities through the general procedure (4.1, 4.2) or through special

types of assessment (4.4–4.6), and could carry out the simple steps described in 4.3.

2. Complex assessments can be built up by iterating these simple steps. For example, the usual process of statistical inference can be broken down into simple steps which involve assessments of prior probabilities and sampling models (4.3.1), refinement to a product space (4.3.3), combination by the conjunction rule (4.3.8), and conditioning (4.3.7) to produce posterior probabilities.

3. Alternatively, \underline{P}_1 can be found directly by maximization over the non-negative reals, $\underline{P}_1(X) = \max \{\underline{P}_0(X - \lambda Z) : \lambda \geq 0\}$. Similarly $\mathcal{E}_1 = \{Y + \lambda Z : Y \in \mathcal{E}_0, \lambda \geq 0\}$. A different method of computing a new model, by applying simple coherence constraints to propagate the effects of the new judgements, is described by Quinlan (1983). The new model obtained by Quinlan's method is less precise than that defined by natural extension, because the coherence constraints he uses are not the strongest possible. Quinlan also suggests a way of modifying the judgements when they incur sure loss.

4. This \underline{P}_5 was computed in finding the extreme points of \mathcal{M}_0 , in 4.2.2.

5. In de Finetti's (1974) approach, evaluation of new random quantities requires, in effect, refinement of Ω . See 2.1.3 and 2.1.4. Manski (1981) analyses the decision problem of when a Bayesian should refine Ω .

6. British football pools do distinguish between the events S and G .

7. Even if the initial model \underline{P}_0 is a linear prevision, the new model will be imprecise unless the multivalued mapping is a coarsening.

8. Here \mathcal{M}_1 is constructed by transferring the probability masses $\underline{P}_0(W \cup L) = 0.6$ to $\{O, M\}$, $\underline{P}_0(D) = 0.25$ to $\{N, M\}$, and the remaining mass 0.15 to Ω . Hence the lower prevision \underline{P}_1 is defined by $\underline{P}_1(X) = 0.6 \min\{X(O), X(M)\} + 0.25 \min\{X(N), X(M)\} + 0.15 \min\{X(N), X(O), X(M)\}$.

9. See 5.13 for discussion and further examples of belief functions. Dempster (1967a) and Shafer (1976) are important references. If the restriction to events of the initial model \underline{P}_0 is a belief function, then so is the new model \underline{P}_1 . More generally, \underline{P}_1 is the lower envelope of the belief functions generated by the linear previsions in \mathcal{M}_0 , but \underline{P}_1 need not be a belief function. (It can be any coherent lower probability.)

10. Precise assessments concerning Ω_0 might be justified if You have extensive frequency information. In the example, You might record the outcomes (W, D or L) of a long series of football games, use the past relative frequencies as precise assessments of probabilities concerning Ω_0 , and generate a belief function on Ω_1 through the multivalued mapping. However, in such cases You will typically have further information about the relative frequency of events in Ω_1 . In the football example, the multivalued mapping generates vacuous probabilities for ‘more than one goal’, $\underline{P}_1(M) = 0$ and $\bar{P}_1(M) = 1$. Even a basic knowledge of football would enable You to sharpen

- these assessments, but then the new model may not be a belief function.
11. In terms of the simplex representation for $\mathcal{P}(\Omega_1)$, \mathcal{M}_1 is the convex hull of \mathcal{M}_0 plus the vertices representing states in Ψ . (Here \mathcal{M}_0 is contained in $\mathcal{P}(\Omega_0)$, which is identified with the subset of the simplex on which $P(\Psi) = 0$.)
 12. This rule can be used to combine many models $\mathcal{E}_1, \dots, \mathcal{E}_n$ by forming the convex hull of their union, or the intersection of $\mathcal{M}_1, \dots, \mathcal{M}_n$, provided this is non-empty. It is not suggested that the conjunction rule, or any other rule of combination, should be used mechanically. Where possible, You should compare and combine the bodies of evidence on which the different models are based.
 13. This is not true for Bayesians, because precise assessments are either identical or inconsistent.
 14. Several kinds of conflict are discussed in 5.2 and 5.4.
 15. Other rules, intermediate between the conjunction and unanimity rules, can be used when a more precise model is needed. For examples and comparison of these rules, see Walley (1982).

Section 4.4

1. There has been a great deal of interest in classificatory probability from philosophers, e.g. Black (1970), Carnap (1962), Day (1961), Kneale (1949) and White (1975), and others concerned with the meaning of 'probable' in ordinary language, e.g. Beyth-Marom (1982) and Kent (1964). There has been little formal or mathematical development, but see Walley and Fine (1979), Burgess (1969), Hamblin (1959), Rescher (1968) and Suppes (1974).
2. This means that the gamble $B - \frac{1}{2}$ is almost desirable. You may judge that both B and B^c are probable. 'Probable' should really be called 'almost probable' to conform to the earlier usage. Of course, if You judge that B is 'probable' in the stronger sense that You are willing to bet on it at even money, then B is 'probable' in the sense used here.
3. These preferences can be distorted if You have additional interest in the occurrence of B or B^c . See Appendix H1.
4. An example of a classification \mathcal{I} that incurs sure loss is given in 2.7.5.
5. When Ω is finite, \mathcal{I}^* consists of all events C such that $C - \frac{1}{2} \geq \sum_{j=1}^n \lambda_j(B_j - \frac{1}{2})$ for some $B_j \in \mathcal{I}$ and $\lambda_j \geq 0$.
6. When \mathcal{I} is coherent, the minimal \underline{P} satisfying (d) of the theorem, and the largest class \mathcal{M} satisfying (e), are the natural extensions of \mathcal{I} . Because the classification \mathcal{I} is equivalent to assessments of lower probabilities, the natural extensions are determined by the lower probabilities of events. That is not true for comparative probability orderings (4.5.2).
7. As is the case with lower probabilities (2.7), these simple properties are strictly weaker than coherence. The listed properties are all consequences

of 4-coherence. (Define n -coherence as in Appendix B4, with every set B_j replaced by A_j^c , and $A_j \geq A_j^c$ just when A_j is probable.) In fact, 1-coherence is equivalent to the first property, 2-coherence to the first two, and 3-coherence to the first three. Walley and Fine (1979) adopt the first two properties, together with completeness, as their basic axioms. A complete classification \mathcal{I}^* that is 3-coherent but not 4-coherent can be constructed from Example 2.7.5. Let \mathcal{I}^* consist of all events A such that $\underline{P}(A) \geq \frac{1}{2}$. Then \mathcal{I}^* contains exactly one event from each pair (A, A^c) . It follows that \mathcal{I}^* is complete and satisfies all the listed properties, except for (v), which fails when $A = B_1, B = B_2, C = B_3$. Thus \mathcal{I}^* is 3-coherent but not 4-coherent. Also \mathcal{I}^* incurs sure loss (since it contains \mathcal{I}). This shows that Theorem 1 of Suppes (1974) is incorrect.

8. When Ω is finite, \mathcal{I} can be represented by an additive probability \underline{P} in 4.4.3(d) if and only if \mathcal{I} is coherent and complete. It is not clear whether this holds for infinite Ω .

Section 4.5

1. Conditional classifications, involving judgements that one event A is probable conditional on another event B , can provide essentially the same information as comparative probability orderings. The behavioural translation of the conditional judgement is that $B(A - \frac{1}{2})$ is almost desirable, which is equivalent to the judgement $A \cap B \geq A^c \cap B$ since $(A \cap B) - (A^c \cap B) = 2B(A - \frac{1}{2})$. The formal properties of conditional classifications are studied by Walley and Fine (1979) and Rescher (1968). An unconditional classification can be regarded as a special type of comparative probability ordering, in which events are compared only with their complements and $A \geq A^c$ if and only if A is probable.
2. There is quite a large literature on comparative probability. See especially the surveys of Fine (1973, 1977a) and Fishburn (1986), and also Carnap (1962), de Finetti (1931), Kaplan and Fine (1977), Keynes (1921), Koopman (1940a, 1940b), Kraft, Pratt and Seidenberg (1959), Krantz *et al.* (1971), Savage (1972a) and Villegas (1964).
3. Equivalently, \geq avoids sure loss if, for any finite set of comparisons $A_j \geq B_j$, there is some state ω such that $\sum_{j=1}^n A_j(\omega) \geq \sum_{j=1}^n B_j(\omega)$. Fine (1977a) uses the terms 'almost additive' for the orderings that avoid sure loss, and 'strictly non-additive' for the others.
4. Note, however, that it is possible to have $A > B$ but $A \approx^* B$, because $B \geq^* A$ but not $B \geq A$. This happens in the example of Appendix B6.
5. When Ω is finite, this condition simplifies to: $\Omega > \emptyset$, and $A \geq B$ whenever $m(A - B) \geq \sum_{j=1}^n (A_j - B_j)$ for some integers $m \geq 1$, $n \geq 0$ and $A_j \geq B_j$. When also the ordering is complete, this condition can be further simplified by assuming $m = 1$.

6. The minimal such \underline{P} is the natural extension of \geqslant . In general, the representation (d) is different from the representations (f) $A \geqslant B$ if and only if $\underline{P}(A) \geq \bar{P}(B)$, and (g) $A \geqslant B$ if and only if $\underline{P}(A \cap B^c) \geq \bar{P}(B \cap A^c)$, although (d) is equivalent to (g) when \underline{P} is 2-monotone. Under (f) and (g), but not under (d) in general, the ordering \geqslant is determined by the upper and lower probabilities of events. Under (d), it is possible to have $A > B$ but $\underline{P}(A) = \underline{P}(B)$ and $\bar{P}(A) = \bar{P}(B)$.

7. De Finetti (1975, Appendix 19) gives a similar example.

8. Most of this work has been based on de Finetti's (1931) axioms for comparative probability, which include completeness. De Finetti's other axioms are consequences of coherence. In fact, his axioms are equivalent to completeness plus 3-coherence; see Appendix B5.

9. This follows from Scott (1964). As in the case of classificatory probability, it is not clear whether this remains true for infinite Ω . All additive orderings \geqslant (those which have an agreeing additive probability measure) are complete and coherent, but I conjecture that there are complete coherent orderings which are not additive. Of course, every coherent ordering has an almost agreeing additive probability measure (4.5.1), but this can have $P(A) = P(B)$ when $A > B$. Armstrong (1987) gives other results on the existence of agreeing probability measures.

10. Strictly, You must compare $A \times \Omega_0$ with $\Omega \times B$, two events in the product space.

11. This is a popular way of constructing additive probabilities. See DeGroot (1970), French (1982), Lindley (1971a), Pratt, Raiffa and Schlaifer (1964), Raiffa (1968), Savage (1972a, pp. 33, 38–9), Spetzler and Stael von Holstein (1975), and Villegas (1964).

12. The issue of comparability is discussed by Fine (1977a, pp. 110–3), Kaplan and Fine (1977), Keynes (1921) and Lindley (1971a, pp. 18–26).

13. These upper and lower probabilities are not necessarily coherent. They should be replaced by their natural extension, or by the natural extension of \geqslant (if other kinds of comparisons are made).

14. This kind of measurement process is suggested by de Finetti (1974, section 5.7) and Suppes (1974). Suppes assumes that a complete ordering \geqslant is defined on a field of events which contains a finite sub-field \mathcal{S} of standard events, whose probabilities are precisely determined. Upper and lower probabilities are measured directly through comparisons with standard events, by $\bar{P}(A) = \min\{P(S): S \in \mathcal{S}, S \geqslant A\}$ and $\underline{P}(A) = \max\{P(S): S \in \mathcal{S}, A \geqslant S\}$. These probabilities are easy to compute, but they may be incoherent, even in the simplest examples, because there is further information in the ordering \geqslant which may yield more precise probabilities for A . An example is given in Walley (1984–85, section 4.6).

Section 4.6

1. Here the lottery is an auxiliary experiment (4.5.7). The special case $\mu = \frac{1}{2}$ represents a judgement that A is probable (4.4).
2. Alternatively, use Appendix A3 and sections 3.1, 3.2.
3. Special cases of this model, in which You assess precise probabilities, are examined by Cano, Hernandez and Moreno (1985), Kudō (1967) and Manski (1981).
4. These assessments were made in the experiment described in Appendix I. Because Ω has only three points, the upper and lower probabilities of all events are determined by those of singletons.
5. Bayesians often elicit additive probabilities through precise assessments of probability ratios or density functions, the case $l = u$. See Berger (1985a, section 3.2).
6. This could be done graphically, by drawing unnormalized upper and lower histograms in which ω_0 has height one unit.
7. When Ω has n states, the assessments determine $2(n - 1)$ hyperplanes, each of which contains $n - 2$ vertices of the simplex.
8. If these density ratios are regarded as probability ratios for small neighbourhoods around ω and ω_0 , the assessments of l and u have the same behavioural interpretation as those in 4.6.2. Of course, the methods of 4.6.1 and 4.6.2 can also be applied to a continuous space by partitioning it into finitely many events.
9. A different model is obtained by allowing the partition in 4.6.1 to become arbitrarily fine. This leads to upper and lower density functions u and l , but \mathcal{M} consists of all *normalized* density functions which lie between u and l .
10. Walley and Pericchi (1988) study a simple example, in which u is constant and l is proportional to a Normal density with mean ω_0 .
11. Intervals of measures were introduced by DeRobertis and Hartigan (1981). See also Hartigan (1983), Walley and Pericchi (1988).
12. This model avoids sure loss provided that l has finite integral with respect to v . DeRobertis and Hartigan (1981) allow l to be an improper density, with infinite integral.
13. More generally, upper and lower previsions are implicitly determined by formulas in DeRobertis and Hartigan (1981).
14. Other types of neighbourhood are suggested by Huber (1981, p. 271), Walley (1982), Wolfson and Fine (1982) and McClure (1984).
15. When an imprecise model \underline{P} is assessed, it may sometimes be useful to add further imprecision by forming a neighbourhood of $\mathcal{M}(\underline{P})$. The constant odds-ratio and linear–vacuous neighbourhoods are especially tractable; see Walley and Pericchi (1988). This avoids the difficulty of assessing precise probabilities P_0 .

16. Huber (1973), Berger (1985a), Berger and Berliner (1986) have studied δ -contamination neighbourhoods. Walley and Pericchi (1988) compare these with constant odds-ratio neighbourhoods, intervals of measures, and other models. The neighbourhood models are relatively difficult to specify, and the inferences they generate depend on the sample size in an unreasonable way. The δ -contamination neighbourhoods have also been used as imprecise sampling models in frequentist studies of robustness, e.g. Huber (1981).
17. It appears that neighbourhoods have been chosen, in the previous studies, to include all ‘reasonable’ or ‘plausible’ prior distributions. This is the sensitivity analysis interpretation criticized in 2.10.4 and 5.9. Other criteria are suggested by Walley and Pericchi (1988).
18. This model is finitely generated, so the natural extensions can be computed by the methods of 4.2.1.
19. If the distribution function F lies between \underline{F} and \bar{F} , and there is a unique value x for which $F(x) = \rho$, then $\underline{x}_\rho \leq x \leq \bar{x}_\rho$. In that sense, \underline{x}_ρ and \bar{x}_ρ can be regarded as lower and upper bounds for the ρ -quantile of X .
20. This model is considered by Cano, Hernandez and Moreno (1985), Kudō (1967) and Manski (1981). Again it is finitely generated.
21. Bayesians need to determine a unique distribution function F . They presumably do so by drawing a curve through the points $F(x_j) = \rho_j$ in some arbitrary way.
22. For the theory of natural-conjugate families, see Raiffa and Schlaifer (1961), DeGroot (1970, Ch. 9), Diaconis and Ylvisaker (1985). Special classes of conjugate priors are studied by Leamer (1982), Pericchi and Walley (1989), Polasek (1984, 1986), Walley and Pericchi (1988).
23. Theorem 7.8.1 shows that this form of statistical inference is coherent. When the sample space is infinite, other coherent inferences may be possible.
24. Some examples of these assessments are given in 5.3 and 5.4.
25. See Berger (1985a, section 4.7.9), and other references therein.
26. However, this approximation, in the weak* topology, is a very crude one. As noted in Appendix D4, any linear prevision can be approximated (in the weak* topology) by a finite mixture of degenerate linear previsions. See Dalal and Hall (1983), Diaconis and Ylvisaker (1985). The conjugate models could be made more flexible by allowing the extreme points of \mathcal{M}_0 to be any finite mixtures of conjugate priors (rather than pure conjugate priors). Such models could be used to approximate any given \mathcal{M}_0 . These classes generate posterior classes of the same form, but the updating is more difficult because the mixing coefficients change.
27. In hypothesis testing, for example, the two hypotheses may be required to have vacuous prior probabilities; see Pericchi and Walley (1989).

Chapter 5

Section 5.1

1. There is considerable variation in terminology in the literature. Our usage of the term uncertainty agrees with some of the economic literature (e.g. Hey, 1979, p. 41), but many economists distinguish uncertainty from *risk*, meaning knowledge of objective probabilities, whereas Bayesians assume that all uncertainty is determinate and can be modelled by precise subjective probabilities (e.g. Harsanyi, 1977). Our distinction between determinate and indeterminate uncertainty follows Knight (1933, p. 46). Indeterminacy is sometimes called ‘true uncertainty’ or ‘strict uncertainty’ in economics, and ‘partial ignorance’ in the statistical literature. Williams (1976) and Levi (1985) refer to ‘indeterminate probabilities’, whereas we treat indeterminacy as a property of beliefs rather than probability models. *Vagueness* might be used instead of indeterminacy, as in Walley (1984–85), but this risks confusion with the Bayesian notion of ‘vague priors’, the different usage of Raiffa (1968), or the vagueness (ambiguity) of ordinary language. (Zadeh, 1978, uses the term ‘imprecision’ for this kind of ambiguity, which is discussed in 5.11.) Our usage of ‘imprecise probabilities’ is consistent with Good (1950, section 7.4) and Levi (1985).
 2. This behavioural definition of certainty corresponds to the notion of practical possibility (2.1.2): You are certain that $\omega = \omega_0$ when no other state is practically possible. Other states may be epistemically possible.
 3. This issue is discussed in 5.6.
 4. Assuming that \underline{P} is defined on a linear space \mathcal{K} and coherent, Δ is a semi-norm on \mathcal{K} . That is, Δ satisfies $\Delta(X) \geq 0$, $\Delta(\lambda X) = |\lambda| \Delta(X)$ for all real λ , and $\Delta(X + Y) \leq \Delta(X) + \Delta(Y)$. Hence $\Delta(X)$ can be interpreted as a distance of X from the linear space containing all gambles with precise previsions, $\{X : \Delta(X) = 0\} = \{X : \bar{P}(X) = \underline{P}(X)\}$. Giles (1976, p. 62) calls $\Delta(X)$ the indefiniteness of X .
- The degree of imprecision could be measured in other ways, e.g. by measures of the form $\rho(X) = \zeta(\bar{P}(X)) - \zeta(\underline{P}(X))$ where ζ is a strictly increasing function. The logit function $\zeta(x) = \log(x/(1-x))$, which transforms the probability interval (0, 1) onto the real line, is natural when considering events. This gives $\rho(A) = \log(\bar{P}(A)\bar{P}(A^c)/\underline{P}(A)\underline{P}(A^c))$. For example, if $\underline{P}(A) = 0.45$, $\bar{P}(A) = 0.55$, $\underline{P}(B) = 0.98$ and $\bar{P}(B) = 0.99$, then A has less precision than B on the Δ -scale, but A has more precision than B on the ρ -scale. These measures were suggested by Good (1962b, 1971a).
5. This terminology is slightly inconsistent, as imprecision is a property of

probability models whereas uncertainty is a property of beliefs. It is perhaps misleading to suggest that there is a single concept of ‘degree of uncertainty’. In general, standard deviation and entropy are two very different measures. (The former measures an average distance of X from its mean, while the latter does not depend at all on the possible values of X , but only on their probabilities.)

6. This is true provided θ , the underlying chance of A , is not zero or one. With chance one, the posterior variance or entropy concerning A will tend to a limit as $n \rightarrow \infty$. This limit may be larger than the value for $n = 1$; that is always the case when $\frac{1}{3} \leq \theta \leq \frac{2}{3}$ and You adopt a uniform prior to reflect ignorance about θ . Of course, the posterior degree of uncertainty concerning θ does tend to zero as $n \rightarrow \infty$, under both Bayesian and imprecise probability models. (For example, the posterior upper and lower standard deviations for θ tend to zero at rate $n^{-1/2}$ under general conditions, whereas the degree of imprecision for θ tends to zero at rate n^{-1} . See Appendix G for examples.) In unusual cases, e.g. when P is the lower envelope of two degenerate priors for θ , it can happen that the lower standard deviation of θ tends to zero, but $\Delta_n(A)$ does not, as $n \rightarrow \infty$. (The upper standard deviation is at least $\frac{1}{2}\Delta_n(A)$, so this cannot tend to zero unless $\Delta_n(A)$ does; see Appendix G4.)

Section 5.2

1. This distinction is discussed by Levi (1985).
2. Under fairly general regularity conditions, as the sample size increases, the posterior degree of imprecision concerning a real statistical parameter is asymptotically proportional to the reciprocal of the observed Fisher information.
3. Multivalued mappings are often useful in modelling this kind of information. See 4.3.5 and 5.13.
4. Other kinds of conflict are possible, e.g. between statistical data obtained from several experiments.
5. This unanimity rule, and other rules for aggregating beliefs, are studied by Walley (1982). Fellner (1965) similarly uses imprecision to model ‘controversial’ probabilities. The extent of disagreement between experts will often be important. If several assessments of accident probabilities for a nuclear reactor have different orders of magnitude, it is misleading to aggregate them into a single precise probability. Disagreement amongst expert opinions is the main source of imprecision in the model studied by Walley and Compello de Souza (1986).
6. Consider the evolution of future oil prices (5.2.4).
7. Some Bayesians seem to regard this as the major source of imprecision; see Berger (1980, p. 29) and Jeffreys (1931, sections 2.5).
8. See Chapters 3 and 8 for examples.

9. This kind of ‘approximate measurement’, resulting in upper and lower bounds on the ‘true’ precise probabilities, has been discussed by Fishburn (1964) and Suppes (1974). See also the discussion of extraneous measurement structures in 4.5.

10. If beliefs are indeterminate but several elicitation methods (e.g. different scoring rules) are used to produce precise probabilities, then it is likely that the elicited models will be inconsistent, because the choices needed to determine a precise model are arbitrary. Winkler (1967a) gives empirical evidence of this. The variation between elicitation methods can be regarded as a type of conceptual imprecision, in the sense of Mellor (1967, 1971), or as a type of conflict between assessment strategies (section 5.2.3). Psychological instability in beliefs is discussed by Fellner (1965), and Savage (1971) reports that it is widely observed in empirical studies of preferences.

11. Leamer (1986) considers other ways in which this can happen.

Section 5.3

1. Karl Pearson (1920, 1921) called this ‘the fundamental problem of practical statistics’. The problem has a long history, including ‘solutions’ by Bayes (1763), Laplace (1812, 1814), Carnap (1952) and Fisher (1956). See also the Bayesian approaches of Lindley and Phillips (1976), Geisser (1984) and Novick (1969).
2. These constraints are necessary and sufficient to ensure that the density is integrable. This is a non-standard parametrization of the beta family. (The usual parameters are st and $s(1-t)$.) See Raiffa and Schlaifer (1961) for properties of the beta–Bernoulli model.
3. It is shown in section 7.8.2 that these posterior probabilities are coherent with the beta–Bernoulli model, and in section 8.4.9 that they are uniquely coherent provided $t_0 > 0$ and $\bar{t}_0 < 1$. The quantity $s_n(\bar{t}_n - t_n) = s_0(\bar{t}_0 - t_0)$ remains constant under sampling. Its constant value S can be regarded as an underlying learning parameter; see 5.3.6.
4. More generally, we might say that You are near-ignorant concerning a gamble X when $\underline{P}(X) = \inf X$ and $\bar{P}(X) = \sup X$. Since X is bounded, this is equivalent to the conditions $\underline{P}(\{\omega: X(\omega) > x\}) = 0$ and $\bar{P}(\{\omega: X(\omega) > x\}) = 1$ whenever $\inf X < x < \sup X$. We use the term ‘near-ignorance’ to indicate that, although Your prior probabilities for A are vacuous, those for θ are not; You are not completely ignorant about the value of θ . The class R_0 studied here is obviously not a unique characterization of near-ignorance, as any larger class will have the same property, e.g. the class R defined in note 6.
5. Coherent assessments of y and z must agree, whatever prior class is chosen.

6. This can be confirmed by examining the larger region of beta priors $R = \{(s, t) : 0 < s < \bar{s}, 0 < t < 1\}$. In cases of very little prior information about θ , the class R (with a large value \bar{s}) may be a more appropriate model than the smaller class R_0 of 5.3.2. However, R generates the same posterior probabilities for A as the class R_0 with $s_0 = \bar{s}$. Since the size (imprecision) of R increases with \bar{s} , smaller values of s_0 or \bar{s} reflect greater prior precision and greater prior information about θ .

This can be reconciled with the accepted wisdom that, for a single beta (s_0, t_0) prior, *larger* values of s_0 reflect more prior information about θ , since they give the prior mean t_0 greater weight in determining the posterior mean t_n . See note 9. Increasing s_0 makes the beta (s_0, t_0) density more concentrated around t_0 , as indicated by the decreasing variance $t_0(1 - t_0)/(s_0 + 1)$, but this produces greater variation amongst these densities as t_0 varies from zero to one, leading to greater imprecision in the near-ignorance (s_0) model.

7. In the limit as $s_0 \rightarrow 0$, the posterior generated by the near-ignorance prior, or by any beta (s_0, t_0) prior, is a precise beta $(n, m/n)$ distribution. The limiting posterior upper and lower probabilities of A are both equal to the observed relative frequency m/n . After observing the outcome of a single trial, Your posterior probability for A is either one, so that You are willing to bet on A at any odds, or zero, so that You are willing to bet against A at any odds. That seems highly unreasonable, even as an approximation, except in the special case where You have prior information that θ is either zero or one. (In that case, a single observation is very informative.) This shows that, in either the Bayesian or near-ignorance models, small values of s_0 reflect strong prior information.

8. See Popper (1959a, Appendix *ix), Keynes (1921, p. 77, 345) and Peirce (1878, p. 179). Peirce suggested that a second number is needed to measure the ‘amount of information’: ‘In short, to express the proper state of our belief, not *one* number but *two* are requisite, the first depending on the inferred probability, the second on the amount of knowledge on which that probability is based.’ The two numbers could be taken to be $\frac{1}{2}(\underline{P}_n(A) + \bar{P}_n(A))$ and $\Delta_n(A)$. For other criticisms of Laplace’s rule of succession, see Keynes (1921, Ch. 30) and Fisher (1956, Ch. 2).

9. Hence $S = s_0(\bar{t}_0 - t_0)$, $M = s_0t_0$ and $N = s_0(1 - \bar{t}_0 + t_0)$. Here M and N can be any real numbers satisfying $0 \leq M \leq N$, not necessarily integers.

The precise beta (s_0, t_0) prior adopted by Bayesians can be regarded as a limit of R_0 as t_0 and \bar{t}_0 both converge to t_0 . Then $S \rightarrow 0$, $N \rightarrow s_0$ and $M \rightarrow s_0t_0$. Thus the beta (s_0, t_0) prior can be regarded as the posterior for a near-ignorance prior with a very small value of S , after observing s_0t_0 successes in s_0 trials. As often claimed by Bayesians, s_0 can be regarded as an ‘equivalent sample size’, and larger values of s_0 reflect greater prior

information. This is based, however, on choosing S to be very small, which already presupposes substantial evidence that θ is very close to zero or one (see note 7). So any beta (s_0, t_0) prior, even one with small s_0 , appears to embody very substantial information about θ . Raiffa and Schlaifer (1961, section 3.3.4) discuss the problems in interpreting s_0 as an ‘equivalent sample size’ or ‘amount of prior information’.

The model can be extended to allow a range of values of s_0 , by defining $R_0 = \{(s_0, t_0) : \underline{S} < s_0 - N < \bar{S}, s_0t_0 > M, s_0(1 - t_0) > N - M\}$, where $\underline{S} \geq 0$, $\bar{S} > \underline{S}$ and $0 \leq M \leq N$. After observing m successes in n trials, the posterior has the same form, with \underline{S} , \bar{S} unchanged and M, N replaced by $M + m, N + n$. This generates the same posterior probabilities for A as the model 5.3.6, but with S replaced by \bar{S} .

10. Alternatively, \bar{t}_0 and t_0 can be directly assessed as the prior upper and lower probability for A .

11. Any such measure will have the desirable property $i(A) = i(A^c)$, since $\Delta(A) = \Delta(A^c)$. This concept of amount of information seems to be related to the ‘amount of knowledge’ discussed by Peirce (1878, p. 178) and the ‘weight of an argument’ or ‘evidential weight’ of Keynes (1921, Ch. 6). These concepts are quite different from Bayesian definitions of ‘weight of evidence’, e.g. Good (1985).

12. The sample information depends on the prior class, but only through the underlying learning parameter S . These results extend to the larger class of beta priors defined in note 9, provided S is replaced by \bar{S} .

13. Several other measures of ‘amount of information’ have been defined, in terms of precise probabilities, by information theorists. See, for example, Shannon and Weaver (1949), Blachman (1966), Kullback and Leibler (1951), Kullback (1959), Good (1966, 1971a, 1975, 1983), Lindley (1956), Amaral-Turkman and Dunsmore (1985), Aczel and Daroczy (1975), and Goel and DeGroot (1981). The sample information provided by Bernoulli observations B concerning A could be measured by: (a) Shannon’s mutual information between A and B , (b) the reduction in entropy (or any other measure of uncertainty) of A on observing B , or (c) the Kullback–Leibler directed divergence between prior and posterior probabilities for A . These are measures of the discrepancy between prior and posterior probabilities for A . None of these measures is additive across independent samples; two independent observations may provide less information than each one does. Indeed, according to the first two measures, observations frequently provide negative information! The directed divergence is always non-negative, but it cannot be interpreted as a difference between prior and posterior measures of information. Under the three measures, the amount of sample information depends strongly on the discrepancy between the prior probability of A and the observed relative frequency, but it is relatively insensitive to sample size.

These properties suggest that the ‘amount of information’ in statistical data can be more adequately measured through degrees of imprecision rather than through precise probabilities.

Our definition of sample information can also be compared with the statistical measure of ‘observed information’ defined by Fisher (1925) and Edwards (1972). This measure is again not additive across independent samples. (The expected Fisher information is additive, but it is a function of the unknown θ .) In large samples, our measure of sample information is approximately proportional to the observed Fisher information for θ . See Machlup and Mansfield (1983) for a survey of other meanings of ‘information’.

Of course, the attractive properties of the measure i for the beta–Bernoulli model will not hold for general statistical models. The posterior degree of imprecision can be increased by prior–data conflict or other sources of imprecision, and then the measure of sample information can be negative. However, the specific model in 5.3.6 and 5.3.7 can be extended to other statistical problems.

For example, in the case of n independent observations from a Normal (θ, σ^2) distribution with known σ^2 , take the prior class \mathcal{M}_0 to consist of Normal (μ, η_0^2) distributions with $\underline{\mu}_0 < \mu < \bar{\mu}_0$ and constant η_0^2 . This class can be specified through the parameters $N = \sigma^2/\eta_0^2$, $M = N(\bar{\mu}_0 + \underline{\mu}_0)/2$ and $S = N(\bar{\mu}_0 - \underline{\mu}_0)$, which have a similar interpretation to 5.3.6. The posterior class then has the same form as the prior class, with S unchanged, and N, M updated to $N + n, M + m$, where m is the sum of the n observations. The results in 5.3.7 carry over to this model.

Section 5.4

1. Compare this model with 5.3.6, defined by fixed s_0 and varying t_0 . The parameters \underline{s}_0 and \bar{s}_0 can be assessed by the methods of 5.3.3, while t_0 can be assessed as a precise betting rate on A . Good (1965, p. 19) uses the model with $t_0 = 0.5$ to represent beliefs about an ordinary coin, and assesses \underline{s}_0 and \bar{s}_0 by considering posterior probabilities based on ‘imaginary results’. It is useful to regard $(\underline{s}_0, \bar{s}_0)$ as a range of weights given to the prior mean t_0 in determining the posterior mean.
2. Unless $r_n = t_0$, the posterior region R_n has a different form from the prior region R_0 .
3. For fixed r_n , $\Delta_n(A)$ is maximized when $n = (\underline{s}_0 \bar{s}_0)^{1/2}$. For this sample size, the prior and data have similar weight in determining the posterior probabilities of A , and the effect of a fixed degree of prior–data conflict is maximized.
4. Carnap (1952) has proposed a wider ‘continuum of inductive methods’,

each of which corresponds to adopting a precise beta $(s_0, 0.5)$ prior for θ , for some positive value of s_0 . These methods are discussed by Carnap (1980), Good (1965, Ch. 5), Jeffrey (1983) and Fine (1973, pp. 195–6).

5. If Your posterior upper and lower probabilities for A after observing one failure are assessed as \bar{z} and z , then $\bar{s}_0 = \bar{z}/(\bar{t}_0 - \bar{z})$ and $\underline{s}_0 = z/(t_0 - z)$.
6. Indeed, the first term is equal to the posterior degree of imprecision for the prior 5.3.6 (with $s_0 = \bar{s}_0$), while the second term is equal to the posterior degree of imprecision for the prior 5.4.1, provided $|r_n - t_0|$ is replaced by the general measure of degree of conflict. For large n , both terms are of order n^{-1} , and $\Delta_n(A) \approx n^{-1}(\bar{s}_0(\bar{t}_0 - t_0) + (\bar{s}_0 - \underline{s}_0)\kappa)$.
7. Luis Pericchi has shown that the model 5.4.3 can be generalized to the one-parameter exponential family. For example, if independent observations are made from a Normal (θ, σ^2) distribution with known σ^2 , then the appropriate class of priors consists of Normal (μ, η^2) distributions with $\underline{\mu} < \mu < \bar{\mu}$ and $\underline{\eta} < \eta < \bar{\eta}$. The posterior degree of imprecision concerning θ then decomposes in an analogous way to 5.4.3, where the degree of prior–data conflict is measured by the distance of the sample mean from the interval $(\underline{\mu}, \bar{\mu})$ of prior means.
8. The proportion of $\Delta_n(A)$ that is due to prior–data conflict increases with n in general, and tends to a limit as $n \rightarrow \infty$. In this example the limit is 0.6 when $r_n = 0$, and 0.43 when $r_n = 0.2$.
9. This issue has been discussed by Cox and Hinkley (1974, p. 384), Leamer (1978, pp. 60–1), and Berger (1985a, section 4.7).
10. In more complicated problems, it is impracticable to consider in advance how You would react to all possible data sets, which might conflict with prior expectations in many ways. It is then sensible to wait until You see the data before trying to make more precise assessments of ‘prior’ probabilities; where extra precision is needed will depend on the data. (For example, assessments of a prior density function for θ should be as precise as possible in the region where the likelihood function is greatest, but it may be unnecessary to make careful assessments outside this region.) Thus the ‘prior’ depends on the data. That is reasonable provided the ‘prior’ probabilities corresponding to the various possible data sets are consistent (i.e., avoid sure loss). (Bayesian priors which depend on the data must be inconsistent.)

Section 5.5

1. The advocates of noninformative priors include Laplace (1812, 1814), Jeffreys (1946, 1983, Ch. 3), Novick and Hall (1965), Jaynes (1968, 1983), Novick (1969), Zellner (1971, 1977), Box and Tiao (1973), Rosenkrantz (1977), Villegas (1977, 1981), Bernardo (1979) and Berger (1985a, section 3.3). For

criticisms, see Ellis (1843), Mill (1843), Boole (1854), Peirce (1878), Venn (1888), Keynes (1921, Ch. 4), Fisher (1956), Fine (1973, Ch. 6), and Cox and Hinkley (1974, section 10.4). According to Jeffreys (1983, p. 117), noninformative priors 'provide a formal way of expressing ignorance of the value of the parameter over the range permitted'. The other advocates seem to share this view.

2. As seen in 5.3, learning is possible from an initial state of near-ignorance.
3. Noninformative priors are widely used by Bayesians in practical problems, e.g. in statistical analyses based on the Normal linear model. See especially Box and Tiao (1973).
4. These principles date back at least to Leibniz in the seventeenth century (see Hacking, 1975). They were extensively used by Laplace, before being discredited by Ellis, Mill, Boole, Venn and others in the last century. See Keynes (1921), Jeffreys (1983, section 1.4), Fisher (1956) and Fine (1973) for further discussion.
5. Provided A is a non-trivial subset of Ω , i.e. both A and A^c are possible.
6. Peirce (1878) and Keynes (1921, Ch. 4) discuss similar examples, and another is given in 2.9.1. It is well known that the principle of indifference can produce different answers to the same problem, according to how the possibilities are listed; see also the examples of Feller (1957, p. 36), Fine (1973, Ch. 6), Jeffreys (1931, sections 2.4, 3.5) and (1983, pp. 128–9). Jeffreys' answers are quite arbitrary. Keynes proposed that Ω be required to consist of 'ultimate and indivisible alternatives' (1921, p. 64), but it is not clear that these ever exist.
7. Consider any non-trivial event A . Partition A^c into n events B_1, \dots, B_n and let $\Omega_n = \{\omega_0, \omega_1, \dots, \omega_n\}$, where the states ω_j are defined by $\{\omega_0\} = A$, $\{\omega_i\} = B_i$. By the embedding principle, $\underline{P}(A)$ does not depend on n . By the symmetry principle, $\underline{P}(B_j) = \underline{P}(A)$ for $1 \leq j \leq n$. By coherence of \underline{P} , $1 = \underline{P}(\Omega_n) \geq \underline{P}(A) + \sum_{j=1}^n \underline{P}(B_j) = (n+1)\underline{P}(A)$. It follows that $\underline{P}(A) = 0$, and $\bar{P}(A) = 1 - \underline{P}(A^c) = 1$ by coherence. Thus \underline{P} is vacuous.
8. See 2.9.1, 3.8.2 and 5.3.7.
9. See Jeffreys (1983, section 3.1), Geisser (1984) and Berger (1985a, section 3.3.4) for discussion of these proposals. Jeffreys (1983, p. 125) concludes that 'we may as well use the uniform distribution', h_1 , but later (p. 184) defends h_2 . Geisser favours h_1 . Compare these proposals with the near-ignorance prior 5.3.2, which we prefer, and the envelope 5.4.2.
10. The Haldane density is subject to many of the criticisms in 5.5.4, because it corresponds to a uniform density for the real parameter $\psi(\theta) = \log(\theta/(1-\theta))$.
11. See the next example or 5.4.2 for evidence of this.
12. This is noted by Box and Tiao (1973, p. 45), Geisser (1984), and Lindley, in the discussion of Bernardo (1979).
13. See Jaynes (1968) and Berger (1985a, section 3.3.2), for example. The

uniform prior seems to have been introduced by Laplace, and was used also by Gauss (1809).

14. All positive multiples of P_0 are equivalent for Bayesians, in that they generate the same inferences and decisions. Because of this, the above argument for translation-invariance, in which P_0 was assumed to be unique, breaks down. See Hartigan (1964) and Berger (1985a, section 3.3.2). So this type of argument gives little support to the uniform density.
15. It is especially strange that Bayesians have weakened the requirement that P_0 be a countably additive probability measure, which is incompatible with translation-invariance, by admitting unnormalized measures. The normalization condition $P(\Omega) = 1$ is a consequence of coherence, but countable additivity is not. There are many finitely additive probabilities which are translation-invariant. These models, or their envelopes defined in 2.9.7 and 3.5.8, seem more reasonable than the improper uniform density, although they are subject to many of the same objections. See also 7.4.9, 7.6.4 and 8.2.9. Although Jeffreys (1931, 1983, p. 119) regards the normalization condition as a 'convention', he has doubts about improper priors and proposes eliminating them by restricting parameter spaces to be bounded (1983, p. 122).
16. Jeffreys (1983, pp. 119–122) recognizes that it is absurd to adopt a uniform prior for a scale parameter σ , because this makes the event $\sigma > \alpha$ infinitely more probable than $\sigma < \alpha$, and 'this is inconsistent with the statement that we know nothing about σ '. But the prior favoured by Jeffreys for scale parameters can be criticized on similar grounds, as it makes the event $0 < \sigma < \alpha$ infinitely more probable than $\alpha < \sigma < \beta$ (where $0 < \alpha < \beta$).
17. Improper priors can also give rise to the 'marginalization paradoxes' discussed by Stone and Dawid (1972), and Dawid, Stone and Zidek (1973), although Jaynes (1980) argues that there is no 'paradox'. See also Sudderth (1980), Stone (1982) and Hartigan (1983).
18. This example is related to 7.4.4. For an extension to many parameters, see Cox and Hinkley (1974, Ex. 2.39) and Berger (1985a, sections 4.7.9, 4.8.2). The best-known example of inadmissibility concerns the estimation of n Normal means, using a uniform density for the multi-dimensional location parameter; see Stein (1956), Cox and Hinkley (1974, section 11.8), Hill (1974) and Berger (1985a, section 4.8.2, 5.4.3). The Bayes estimator is inadmissible when $n \geq 3$. In both these examples, the quadratic utility functions are unbounded and unrealistic. Incoherence is a more fundamental defect than inadmissibility, because the latter depends on a specific utility function.
19. For example, You can be almost certain that θ^{-1} lies in an arbitrarily small interval around zero, which seems incompatible with complete ignorance. Similarly, the uniform density for θ transforms to the improper Haldane density on $(0, 1)$ for $\psi(\theta) = \exp(\theta)/(1 + \exp(\theta))$, which seems to

model beliefs that ψ is very close to zero or one. See also Fine (1973), Rosenkrantz (1977) and Jaynes (1983) for discussion of Bertrand's paradox, and Fisher (1956, p. 16), Box and Tiao (1973, p. 24), and Cox and Hinkley (1974, p. 377). The vacuous probabilities are invariant under arbitrary transformations of θ .

20. See Box and Tiao (1973, pp. 41–6) and Bernardo (1979). For Bernardo, and Zellner (1977), the noninformative prior depends not only on the sampling model, but also on the specific parameter in which You are interested. This can lead to incoherence when You make inferences about several parameters, even when Θ is finite so that all priors are proper.

21. Or θ may be a location parameter for one experiment, while $\psi(\theta)$ is a location or scale parameter for another experiment.

22. Jeffreys (1983, pp. 138, 182), Box and Tiao (1973, section 1.3.6), and Berger (1985a, section 3.3.3).

23. See Berger (1985a, section 4.6.2), who suggests replacing h_2 by a uniform prior density for (μ, σ) . A. F. S. Mitchell, in the discussion of Good (1967), and Cox and Hinkley (1974, Ex. 10.6) describe an exponential regression model for which a noninformative prior generates improper posteriors, again irrespective of the observations. If the sample space is finite then an improper prior must generate an improper posterior for some possible samples.

24. See Lindley (1965), Stone (1982), Box and Tiao (1973, pp. 20–1). Whether the approximation is adequate will typically depend on the statistical data; see Hill (1974) and Dickey (1976).

25. This example has been discussed by many authors, including Jeffreys (1983, section 5.02), Lindley (1957), Edwards, Lindman and Savage (1963), DeGroot (1970, section 11.6), Cox and Hinkley (1974, Ex. 10.12), Shafer (1982b), Good (1983) and Berger (1985a, section 4.3.3). Similar results are obtained if H_0 is replaced by the hypothesis that θ lies in a bounded set.

26. Box and Tiao (1973, p. 23).

27. Bernardo (1979, p. 114). Box and Tiao (1973, p. 22) hold a similar view.

28. This is a loose statement of the principle of stable estimation, discussed by Jeffreys (1931, section 3.4), von Mises (1942), Good (1950, p. 77), Blackwell and Dubins (1962), Edwards, Lindman and Savage (1963), Lindley (1965), DeGroot (1970), Hartigan (1983) and Berger (1985a).

29. These answers can be challenged on either a frequentist or a Bayesian interpretation; see 7.4 and 7.5 for examples.

30. For some work in this direction, in the case where θ is a real location parameter, see Walley and Pericchi (1988).

Section 5.6

1. See de Finetti (1974, section 3.1.4) and the comments of Lindley on Giron and Rios (1980).

2. The Bayesian may have a preference between the two actions, but it is not determined by his probability and utility assessments. (As suggested in 3.8.6, not all preferences can be modelled in terms of expected utility.) We would also be indecisive in this problem, because the vacuous probabilities do not determine any preference between the actions. For us, the lack of preference reflects the lack of information about Ω , rather than a precise judgement of equality. For Bayesians, the two actions are 'equally good', in that an arbitrarily small change in probabilities or utilities is enough to determine a preference. In our analysis the two actions are incomparable and they remain so when the probabilities or utilities are modified. How would Bayesians choose between the two actions if there were also no information about their consequences?

3. A preference for this option implies only that $\{\omega_1\}$ is probable. More generally, a judgement that one action is optimal, or even a complete preference ordering on actions, will generate probabilities that are imprecise.

4. On this distinction, see Sen (1982, esp. Essay 2). The concept of revealed preference in economics corresponds to choice rather than preference.

5. On this view, which is suggested by Savage (1971, 1972a), decision analysis resembles psychoanalysis: both aim to discover Your unconscious preferences through careful analysis of Your actions, thoughts and feelings. There may be little there to discover, and the analysis may produce little more than a rationalization of arbitrary choices. The view that You must have underlying preferences, or beliefs and values that determine them, seems to leave little scope for freedom of choice or creativity; these could only be displayed by acting irrationally, contrary to Your true preferences. See Kyburg (1978), Suppes (1984, p. 209), and Dennett (1978, Ch. 15) and (1984).

6. This is desirable even in practical decision problems such as medical diagnosis, as recognized by Savage (1972b, p. 133), and Pereira and Pericchi (1988). Walley and Campello de Souza (1986) give a detailed analysis of a practical decision concerning investment in solar heating. The analysis gives considerable insight into the problem, although it results in indecision; the alternative investments are both reasonable, on current information.

7. If, for example, a uniform distribution is used to represent ignorance, as in 5.5.1, the resulting decision will depend on the choice of a possibility space.

8. This seems to be accepted by Savage (1972b, p. 133), Good (1952, 1983, p. 9) and Berger (1984, pp. 142–4).

9. According to the Bayesian theory, precise probabilities and utilities can be assessed whatever Ω is chosen, and these determine a best action. So it is difficult for Bayesians to explain why some choices of Ω are better than others. From our point of view, one choice of Ω is better than another when the first leads to a smaller set of maximal actions than the second. Note also that a Bayesian analysis may be indecisive, if several choices of Ω lead to different actions.

10. This may introduce extra indecision because a_0 cannot be compared with the other actions, especially as the consequences of suspending judgement will often be unclear. Levi (1980, pp. 214–15) advocates suspending judgement whenever that is feasible. It is certainly wise to suspend judgement in inference problems, when conclusions are indeterminate. In statistical hypothesis testing, it may be reasonable, when the evidence is indecisive, to suspend judgement between the hypotheses. In medical diagnosis it may be reasonable to postpone a diagnosis until further tests can be carried out.

11. Some critics of upper and lower probability have confused indecision with suspension of judgement. Thus they translate indecision (lack of preference) between a_1 and a_2 into a preference for a_0 over each of a_1 and a_2 . This is a misunderstanding. Suppose that You have no preference between a_1 and a_2 . If suspending judgement (a_0) is feasible, it may be better than a_1 and a_2 , or it may be worse, or there may be indecision between all three actions. If a_0 is not feasible, You can still choose between a_1 and a_2 . Indecision need not lead to dithering or paralysis; compare with the amusing discussion of Buridan's ass in Sen (1982, Essay 2).

12. Suggested by Simon (1955).

13. This rule is advocated by Gardenfors and Sahlin (1982). Alternatively, one could choose X to maximize $P(X - Z)$, where $Z(\omega) = \max\{X(\omega): X \in \mathcal{X}\}$ is the best possible reward when ω is the true state. Then X minimizes the maximum expected regret ($Z - X$) rather than the maximum expected loss ($-X$). The minimax regret rule is often preferred in statistical decision problems.

14. These decision rules are often called *gamma-minimax* rules. There is a large statistical literature on such rules, including Berger (1984, 1985a, section 4.7.6), Blum and Rosenblatt (1967) and Menges (1966); Berger gives other references. When the prior P is vacuous, gamma-minimaxity coincides with the classical notion of minimaxity. Other ‘robust’ types of statistical decision rule are described by Berger (1984, 1985a).

15. Suppose You assess $P(A) = 0.3$ and $\bar{P}(A) = 0.7$, and You are offered the gamble $X = A - 0.4$. You have no preference between accepting and rejecting X , but the P -minimax rule leads to rejection since $P(X)$ is negative. Similarly, You would reject $Y = 0.6 - A$ by this rule. But if both X and Y are offered You should obviously accept them both, since $X + Y = 0.2$ is a sure gain. (Buehler, 1976, gives a similar example.) This shows that the minimax rule can generate inferior decisions if it is applied to several related decision problems. As far as practicable, the separate decisions should be combined into a single problem before applying the rule. (The P -minimax action in the joint problem is to accept both gambles.) Ellsberg's paradox (1961) is another example of this: the rewards can be modified slightly so that the minimax rule is to choose gamble U over V and X over Y , but (V, Y) together are surely better than (U, X) .

16. The maximality theorem 3.9.5 shows that, if utilities are already precise, all maximal actions can be obtained in this way. The precise probability model might be chosen by assessing a second-order distribution (5.10), or by maximizing entropy (5.12). Note, however, that the decisions which result from definite rules such as maximizing entropy or assigning a ‘noninformative’ second-order distribution will depend on the arbitrary choice of a possibility space Ω . Smith (1965, p. 478) suggests using the same precise probability and utility models for all later decisions, to ensure coherence across different decision problems.

Section 5.7

1. Surveys of these axioms, not restricted to the axioms of precision, can be found in Fine (1973, Ch. 8) and Fishburn (1981, 1986).
2. If Bayesians really can assess fair prices for all gambles, one would expect money to be changing hands amongst them at a great rate. Any two Bayesians should be able to arrange a mutually acceptable bet concerning any event, unless their fair betting rates happen to be exactly equal. (The problem of different utility scales can be avoided by using probability currency (2.2), i.e. bets could be paid off with lottery tickets instead of money.) See also the excellent discussion of Kyburg (1978, esp. pp. 163–4).
3. De Finetti (1974, section 3.3.4). Here de Finetti appears to identify $P(X)$ with Your choice of \bar{x} , but a careful reading suggests that his definition is not as ‘operational’ as it first appears. De Finetti continues by claiming that it is in Your interest to honestly report Your real ‘fair price’ for X ; but that is mistaken, for the reasons noted in Appendix H and by de Finetti himself (1974, p. 93). Moreover, he implicitly assumes in 3.3.7 that You have underlying preferences which fully determine Your responses \bar{x} to his operational measurement procedures. (This assumption, which is more clearly stated on p. 73, is needed to show that the two operational ‘definitions’ of previsions lead to the same measure.)
4. See de Finetti (1975, Appendix 19.3), and also (1937, p. 150), (1974, pp. 193, 198) and de Finetti and Savage (1962). In (1977a, p. 5), he admits that he might be unable to assess a precise probability for a hypothesis because ‘I lack any grounds for an opinion’.
5. Ramsey's (1926) essay is one of the first and most important contributions to a behavioural theory of probability. The essay is tentative, and Ramsey later expressed some doubts about his interpretation of partial belief as a ‘psychological phenomenon’ (1931, pp. 256–7). See also his ‘Further considerations’ (1931, pp. 199–211).
6. Raiffa (1968, Ch. 5, esp. pp. 110–14) defines probability in a similar way to de Finetti and Ramsey, by assuming that there is a fair betting rate $P(A)$

such that You are indifferent between the gamble A and the constant gain $P(A)$. Raiffa argues (p. 109) that You can evaluate $P(A)$ by comparing A with drawings from a calibrating urn. He admits that there will be a 'range of indecision' in Your evaluations, from 0.75 to 0.85 in the example he discusses, but he suggests adopting the 'central value' 0.8 on the grounds that 'we need not be unduly concerned about such critically fine distinctions' (p. 109). Raiffa then proceeds to ignore the imprecision in this evaluation, and concludes (without any justification) that 'The vagueness you feel about the 0.8 measurement does not enter into the analysis.' (p. 120). He discusses (p. 112) some experimental results, based on Ellsberg (1961), which suggest that this vagueness does influence decisions.

7. References are given in 1.8.5, 4.5, Fine (1973, 1977a, 1983) and Fishburn (1986).

8. Jeffreys (1931, section 11.2) discusses Keynes' work, but instead of justifying his assumption of completeness he merely asserts 'It seems to me that all probabilities actually are comparable and that Keynes is merely creating difficulties... But my real objection to Keynes's postulate is that it is one of those attempts at generality that in practice lead only to vagueness.'

9. Jeffreys seems to say (1983, pp. 33–4) that, when p is a tautology, any two propositions q and r are equally probable!

10. Lindley (1971a, p. 21). It is not entirely clear what Lindley means here. He claims that different kinds of numerical probability would be needed for the two types of event, but dismisses this on the grounds that 'the rules of probability have never been seriously challenged'.

11. Additive probabilities avoid sure loss, but they also avoid sure gain; see 2.8.2 and 2.9.3.

12. Keynes (1921, Ch. 3) used the same examples (bookmakers' odds and insurance premiums) to argue that probabilities are *not* precisely determined! Russell (1948, pp. 358–60) argues that the probability You assess when considering insurance will be 'quite vague'. See Einhorn and Hogarth (1985, 1986) and Hogarth and Kunreuther (1985) for evidence of indeterminacy in the insurance problem.

13. A similar example is discussed by Berger (1984, p. 112) and (1985a, p. 285).

14. For experimental evidence that vagueness can influence decisions, see Ellsberg (1961), Einhorn and Hogarth (1985, 1986), and the other references given in section 1.8.7. The nuclear issue is considered by Levi (1980, Appendix), Elster (1983a), and Vasey and Rasmuson (1984).

15. Similarly DeGroot (1970, Ch. 6), following Villegas (1964), bases his system on de Finetti's (1931) axioms. DeGroot offers no justification for the completeness axiom. He requires also that the ordering can be extended to a complete joint order with a uniform distribution on the unit interval. (As

argued in 4.5.7, this is a very strong requirement.) The precise probability of any event A can then be evaluated as the real number p such that the interval $[0, p]$ and event A are equally probable. (This is also required by axiom 2b of Pratt, Raiffa and Schlaifer, 1964.) This shows rather starkly the Bayesian assumption that probabilities can be measured with arbitrary precision.

16. Savage's axiom P2 is another completeness axiom, objectionable in the same way as P1. Fishburn (1975) indicates how Savage's axioms could be modified to accommodate imprecision; his conditions seem to require avoiding sure loss but not coherence. For discussions of Savage's other axioms, see Fishburn (1970, 1975), Fine (1973) and Suppes (1974). Most of the criticisms of Savage's system have focused on his strong partitioning axiom P6, his assumption of independence between states and actions, and the large set of acts (including all constant acts) which must be compared. Other theories of subjective expected utility are surveyed by Fishburn (1970, 1981) and Fine (1973, Ch. 8). These theories, e.g. Anscombe and Aumann (1963) or Jeffrey (1983, Ch. 9), include an axiom that requires completeness of a preference ordering for acts or randomized acts or randomized consequences or propositions, similar in effect to Savage's P1. Aumann (1962) criticizes this kind of completeness axiom and outlines a theory of utility which does not require completeness.

17. See also Savage's comments on partial preference orderings and convex sets of probability measures (1972a, pp. 21, 58), and other discussions of precision in (1972a, section 9.7), (1967, p. 308), (1971, p. 788), (1972b, p. 133), (1977, pp. 7, 12), and Edwards, Lindman and Savage (1963, pp. 201, 208).

18. For discussion of this rule, see de Finetti (1974, section 5.4, 5.5) and Winkler (1967b, 1969, 1971). De Finetti gives a more general scoring rule, involving positive proportionality constants k_j , but that does not affect the argument. A general class of scoring rules is studied by Savage (1971); see also Appendix H8 and 5.12.3(e). Some very general mathematical results are given by Lindley (1982) with a lively discussion of their implications concerning the need for precision.

19. Verify that the difference in penalties $S(x') - S(x)$ is $3\delta^2$.

20. De Finetti (1974, section 3.3.6), Savage (1971, section 8) and Lindley (1982) propose that scoring rules be used to define probabilities operationally. That is, Your probabilities are to be identified with Your 'response' x to the 'stimulus' of a particular scoring rule. If this is done, the previous argument establishes that the probabilities (i.e. responses) should be additive. But probabilities cannot be adequately defined in terms of choices in a specific decision problem, because the choices may be arbitrary and then there is no reason for the same probabilities to apply in other decision problems. Scoring rules are useful in measuring probabilities only

when there are underlying determinate beliefs which determine Your response to the rules, as well as Your decisions in other problems. When Your beliefs are indeterminate, various responses will be reasonable, and it can be reasonable to choose different responses to different types of scoring rules, e.g. logarithmic rather than quadratic. (This is illustrated in 5.12.3(e).)

21. If no events have probability $\frac{1}{2}$, one classificatory response is uniquely optimal under P .

22. Consider a deterministic scoring rule: You must choose some state ω_0 from Ω , and You will obtain reward 1 if ω_0 occurs, zero otherwise. De Finetti (1974, sections 3.1.2, 3.7.3) rightly condemns this kind of deterministic rule, because it does not allow You to express Your uncertainty about Ω . Bayesian scoring rules should be condemned on similar grounds: they do not allow You to express the indeterminacy in Your beliefs about Ω .

23. A different type of argument is given by Cox (1961). See also Cox (1946), Reichenbach (1949), Good (1950, p. 105), Aczel (1966, pp. 319–24) and Tribus (1969, pp. 13–14). Cox assumes (p. 13) that every event has a precise real-valued probability, and that some kinds of probabilities are functions of others. (Specifically, he assumes that $P(A \cap B|C)$ is a function of $P(A|C)$ and $P(B|A \cap C)$, and that $P(A^c|B)$ is a function of $P(A|B)$.) No justification is given for these assumptions. After imposing some regularity conditions and assuming that an arbitrary real function is the identity function, Cox obtains the standard axioms for finitely additive probability. In the following discussion, Cox acknowledges that his assumption of precision is unjustified, for it is only in ‘exceptional’ cases, such as games of chance and statistical randomization, ‘that probabilities can be precisely estimated’. When ‘nothing is known’ about an event, its probability is ‘entirely undefined’ (1961, p. 32).

Section 5.8

1. See Suppes (1974) and Lindley, Tversky and Brown (1979) for the comparison with Euclidean geometry, Ramsey (1926), Mellor (1980) and Lindley (1983) for Newtonian mechanics.

2. This point is made by Savage (1972a, section 4.2).

3. Discussed by Suppes (1974) and Mellor (1971).

4. Mellor (1971) argues that probabilities are not fundamentally different from other physical concepts in this respect; all have ‘conceptual imprecision’ as well as ‘operational imprecision’. But, as noted in (4), there does seem to be a difference in degree and practical importance between the (conceptual) imprecision in probabilities and that in physical quantities such as distances or time intervals.

5. Compare with the ‘idealization’ that the Hundred Years’ War between

England and France occurred on 25 August 1403. This is a poor idealization because of its precision, not because of any ‘measurement error’.

6. Good (1962a) raises this issue.

7. If beliefs about A are indeterminate, it may be ‘easier’ in some sense to assess a precise $P(A)$ than to assess an exhaustive model $\underline{P}(A)$, $\bar{P}(A)$, but the precise probability will not be a correct description of beliefs. Compare the precise, exhaustive and incomplete probability judgements with the judgements that the Hundred Years War occurred (a) in 1403 (precise), (b) between 1338 and 1453 (exhaustive), or (c) between 1300 and 1500 (imprecise but not exhaustive). It is easier to make the first and third assessments than to make the second, but the first is simply wrong, whereas the third is correct but less informative than the second. It is not possible to give exact bounds like (b) for, say, the Italian Renaissance, whose duration is not sharply defined, but we could still give upper and lower bounds like (c).

8. However, the choice of a model may be less crucial when incomplete models are admitted. If a model P is adequate, then there may be many less precise models that are also adequate.

9. Coherence requires $\underline{P}(\emptyset) = 0$ and $\bar{P}(\Omega) = 1$, but imposes only inequality constraints on the lower probabilities of other events. Coherent lower revisions correspond to closed convex sets of linear previsions, which may have infinitely many extreme points.

10. For example, there are computational difficulties in checking whether a specified lower prevision is coherent, and in computing marginal or posterior probabilities or other types of natural extension. There are similar difficulties in Bayesian analysis. They can be overcome by developing general types of models which are coherent and tractable, together with appropriate numerical methods and approximations. Some tractable types of imprecise models are described in sections 2.9, 4.6, 5.3, 7.8, DeRobertis and Hartigan (1981), and Walley and Pericchi (1988).

11. See especially Lindley (1971b) and Berger (1984, 1985a).

12. For example, the ‘most powerful’ Neyman–Pearson tests of simple hypotheses are essentially the Bayesian tests, and the choice of the ‘size’ of the test is equivalent to the choice of a prior probability for the null hypothesis. See Berger (1985a, Ch. 8).

Section 5.9

1. Some other ways of modelling indeterminacy have been suggested by Einhorn and Hogarth (1985), Fellner (1965), Gardenfors and Sahlin (1982, 1983), Kmietowicz and Pearman (1981), Kyburg (1961, 1970, 1974a), Levi

(1974, 1980, 1982, 1985), Nau (1981, 1986), and Shortliffe and Buchanan (1975).

2. Berger (1984) gives a clear account of the basic ideas of Bayesian sensitivity analysis. Other general discussions include Berger (1985a), Dempster (1975), Dickey (1973), Edwards, Lindman and Savage (1963), Fisher (1957), Good (1965, 1983), Hartigan (1983), Kudo (1967), Leamer (1978), Lindley (1971b, Ch. 7), von Mises (1942), and von Winterfeldt and Edwards (1986, Ch. 11). For specific kinds of models, see Berger (1985a, section 4.7), Berger and Berliner (1986), Blum and Rosenblatt (1967), Cano, Hernandez and Moreno (1985), Chamberlain and Leamer (1976), DeRobertis and Hartigan (1981), Dickey (1974), Fishburn (1964, 1965), Fishburn, Murphy and Isaacs (1968), Giron and Rios (1980), Goldstein (1974), Hartigan (1969), Huber (1973), Isaacs (1963), Kadane and Chuang (1978), Leamer (1982), Manski (1981), Pierce and Folks (1969), Polasek (1984, 1986), Potter and Anderson (1980, 1983), Walley and Pericchi (1988), Watson (1974), and other references in Berger (1984, 1985a). Mosteller and Wallace (1964) is an excellent example of practical sensitivity analysis.

3. See Edwards (1961, 1968), Einhorn and Hogarth (1981, 1985), Fellner (1965), Hogarth (1975, 1980), Kahneman, Slovic and Tversky (1982), Pitz and Sachs (1984), Schoemaker (1982), Slovic and Lichtenstein (1971, 1983), Winkler (1967a), von Winterfeldt and Edwards (1986), and other references in 1.8.7. Other versions of the descriptive interpretation are given by Lindley, Tversky and Brown (1979, p. 149), for whom P_T represents an underlying ‘set of coherent probabilities that are distorted in the elicitation process’, Dickey (1980) and Mellor (1980), for whom P_T models Your unconscious beliefs.

4. See also Good (1965, 1975, 1976, 1983) and Levi (1980, 1985). Good’s extensive writings on probability and inference are a fertile source of ideas, most of which are consistent with our approach.

5. Strangely, this translation is inconsistent with a ‘rule of application’ proposed by Good (1962a), under which the judgement ‘*A* is more probable than *B*’ corresponds to the inequality $\underline{P}(A) > \bar{P}(B)$.

6. Good (1962a, p. 323) assumes ‘that a probability *inside the black box* was numerical and precise’. There is a similar statement in Good’s discussion of Smith (1961).

7. Good (1965, p. 35) and (1983, p. 98).

8. Good (1983, p. 123); see also (1965, p. 10).

9. Berger and Wolpert (1984, p. 136). A similar interpretation is suggested by Brown, in the discussion of Lindley, Tversky and Brown (1979).

10. Berger (1984, p. 72). See also Berger (1980, p. 29).

11. Consider the analogy between probabilities and occurrence times. You might determine, after brief reflection, that the Hundred Years War took

place in the period between 1300 and 1500. It would be absurd, however, to claim that only ‘lack of time’ prevents You from ascertaining the precise year in which it took place!

12. See Walley and Pericchi (1988) for other desiderata (properties of \underline{P} and \bar{P}) that seem appropriate when there is little prior information, and for some models which satisfy them.

13. For examples, see Chapter 9, sections 5.3, 5.4, Pericchi and Walley (1989), Walley and Pericchi (1988).

14. Our approach is apparently even closer to that of Levi (1980, 1985), who represents rational beliefs through convex sets \mathcal{M} of additive probabilities but unequivocally rejects a sensitivity analysis interpretation. See 1.8.8.

Section 5.10

1. This approach is suggested by Good (1952, 1962a, 1965, 1980, 1983), Dickey (1980), Mellor (1980), Skyrms (1980), and Lindley, Tversky and Brown (1979). For criticisms, see de Finetti (1972, pp. 145, 189–193), Savage (1972a, Ch. 4) and Levi (1980, Ch. 9). Second-order probabilities are also called ‘type 2 probabilities’ (by Good) and ‘beliefs about beliefs’ (Dickey, Mellor and Skyrms). Higher-order models can be constructed by assessing n th order probabilities concerning the possible $(n - 1)$ th order distributions. There is the possibility of an infinite regress, but this can be terminated by assuming that the n th order probabilities are either degenerate or vacuous when $n > N$.

2. In fact, P_1 is the unique first-order prevision that is coherent with P_2 and the hypotheses in \mathcal{M} . See 6.7.3. A numerical example is given in 5.12.4(b).

3. A similar argument is given by Savage (1972a, Ch. 4) and Levi (1980, pp. 186–9) and (1982), who argues that the model is self-contradictory.

4. This is not to deny that You can be uncertain about Your current behavioural dispositions. To be coherent, however, Your second-order beliefs about these dispositions must be vacuous on the set \mathcal{M} . It is inconsistent to adopt precise second-order probabilities P_2 concerning Your current dispositions, unless P_2 is actually degenerate at P_1 . Mellor (1980) avoids the inconsistency by distinguishing conscious judgements (P_2 and P_1) from unconscious dispositions (P_T), although it seems unrealistic to assume that the latter are precise. A similar distinction is made by de Sousa (1971).

5. The models of Lindley, Tversky and Brown (1979) and Dickey (1980) can be interpreted in this way. The statistical data x are the judgements You make during elicitation, which depend on Your true probabilities P_T through an explicit sampling model. An external observer could use the data (x) to update his prior (P_2) concerning the unknown P_T . It is very difficult in practice to construct a realistic sampling model and prior.

6. This interpretation is considered by Good (1962a, 1965, 1980). He assumes P_T is precise, but allows P_2 to be imprecise.
7. There is an extensive literature on hierarchical models, including Good (1965, 1980), Box and Tiao (1973), Lindley and Smith (1972), Pearl (1988), Pericchi and Nazaret (1987), Schum and Martin (1982), von Winterfeldt and Edwards (1986, Ch. 6), Zellner (1971), and Berger (1985a, sections 3.6, 4.6), who gives further references.
8. See especially 5.5.4(h), and also Pericchi and Nazaret (1987). These dangers are present even when the hyperparameter has small dimension.
9. Because of the difficulties in defining P_T and the mathematical complexity of the space $\mathcal{L}(\mathcal{M})$, second-order probabilities may be more difficult to assess, and less precise, than first-order ones. That is the view of Good (1952, 1983).
10. Suppose, for example, that Your analysis of evidence results in model P_0 . You assess precise probability $1 - \delta$ that the analysis is correct, but in case of error You have no idea about the correct model. Then P_1 is a linear-vacuous mixture (2.9.2).

Section 5.11

1. We can distinguish the following approaches:
 - (a) Fuzzy decision analysis: see Watson, Weiss and Donnell (1979), Freeling (1980), Dubois and Prade (1979, 1980), Jain (1976), and Kickert (1978).
 - (b) Decision theory with fuzzy descriptions of events, actions or consequences: Jain (1976), Tanaka, Okuda and Asai (1976), Stallings (1977), Kickert (1978), Dubois and Prade (1979), and Feagans and Biller (1980).
 - (c) Optimization under fuzzy constraints (without uncertainty): Bellman and Zadeh (1970), and Kickert (1978).
 - (d) Fuzzy logic and approximate reasoning: Zadeh (1973, 1975, 1983), Gaines (1976, 1978), Forsyth (1984), and Prade (1985).

Kickert (1978) surveys various decision theories which are based on fuzzy sets. Feagans and Biller (1980) compare fuzzy models with upper and lower probabilities in the analysis of public health risks. Our discussion concentrates on (a), but many of our criticisms apply also to the other approaches. There is remarkably little attention in this literature to fundamental issues of meaning and assessment.

2. The approaches of Gardenfors and Sahlin (1982) and Nau (1986) resemble fuzzy decision analysis. Instead of membership functions, these two theories use ‘reliability measures’ and ‘confidence weights’ to reflect second-order beliefs.

3. This kind of vagueness is discussed by Black (1970).
4. Numerical translations of ‘probable’ are considered by Beyth-Marom (1982), Kent (1964), Walley and Fine (1979), Wallsten *et al.* (1986), Budescu *et al.* (1988).
5. Watson *et al.* (1979, p. 7) admit that ‘the precise shape of these distributions is somewhat arbitrary’, but claim that ‘it is the general shape of the distributions that matter’.
6. In the special case where μ_p and μ_u are indicator functions, so is μ_a , and then fuzzy decision analysis apparently reduces to Bayesian sensitivity analysis. It appears also that a fuzzy analysis may result in indecision, although the behavioural meaning of the functions μ_a is unclear. The ability of fuzzy decision analysis to model indeterminacy and indecision is an advantage over second-order probabilities if these are assumed to be precise, but not if imprecision is allowed.
7. Compare with Zadeh (1978, 1983).
8. See Gaines (1975) and Freeling (1980) for further discussion.
9. Gaines (1975, 1978), Hersh and Caramazza (1976), Schefe (1980), Giles (1988) and Hisdal (1988) suggest several different probabilistic interpretations of membership functions. It may be more feasible to interpret μ_p as some kind of upper probability; see Prade (1985) and Giles (1988). Giles interprets degrees of membership, in general, as standardized utilities.
10. There is little guidance on assessment in the literature, but see Gaines (1976, section 9), Kickert (1978, Epilogue), Hersh and Caramazza (1976), Watson *et al.* (1979), Freeling (1980), and Wallsten *et al.* (1986). Freeling (1980, p. 344) acknowledges that, in practice, we cannot make the extra assessments needed for a fuzzy decision analysis.
11. See Wallsten *et al.* (1986) for evidence of large interpersonal differences in the meanings of ambiguous probability judgements.
12. Compare with our justification of coherence and the rules for natural extension in Chapters 7 and 8, or with de Finetti’s (1974) justification of Bayes’ rule and the calculus of finitely additive probability, which both rely on a behavioural interpretation.
13. There is considerable debate in the literature about whether the rules are appropriate. (They are obviously inappropriate in some cases, e.g. when A and B are disjoint, or when $A \cup B = \Omega$.) See Bellman and Giertz (1973), Gaines (1975, 1976), Hersh and Caramazza (1976), Zadeh (1978), Thole, Zimmerman and Zysno (1979), Schefe (1980), Giles (1988) and Hisdal (1988). The theories of Shackle (1961) and Cohen (1977), and the expert system MYCIN, use similar rules to combine measures of uncertainty.
14. The simpler structure of upper and lower previsions (or, equivalently, of \mathcal{M}) seems conceptually clearer, more realistic as a description of imprecise reasoning, and much better justified, through coherence and the behavioural

interpretation. See also the conclusions of Kickert (1978, pp. 157–8) and Freeling (1980).

Section 5.12

1. Jaynes has forcefully argued for the PME in many papers, the most important of which are collected in Jaynes (1983). Maximum entropy is also supported by Good (1963, 1965), Tribus (1969), Rosenkrantz (1977), Levine and Tribus (1979), Bernardo (1979) and Berger (1985a, section 3.4). For criticisms, see Fine (1973, Ch. 6) and Dias and Shimony (1981). The PME is extended to a principle for updating probabilities by Williams (1980), and Shore and Johnson (1980). In the maximum entropy approach, as in sensitivity analysis, \mathcal{M} is determined through constraints on linear previsions which represent ‘partial information’ about the true prevision; see Jaynes’ (1968) definition of ‘testable information’.

2. This defines a unique measure P_0 , provided \mathcal{M} is closed and convex (so that it corresponds to a coherent lower prevision \underline{P}), because the function H is bounded above (at least when Ω is finite), continuous and strictly concave. The case of infinite Ω is mentioned in 5.12.4.

3. The ‘reference priors’ of Bernardo (1979) agree with the maximum entropy priors when Ω is a finite parameter space.

4. Tribus (1969, Ch. 5) gives many other examples.

5. See Jaynes (1957, 1968), Berger (1985a, section 3.4) and Tribus (1969) for these descriptions.

6. An alternative rule for selecting P_0 is to minimize the sum of squared probabilities; see 5.12.3(e). Another rule, applicable only in statistical problems, is the ML-2 (type-2 maximum likelihood) rule discussed by Good (1965, 1983) and Berger (1985a, sections 3.5.4, 4.7.2). The ML-2 prior is chosen from \mathcal{M} to maximize the prior predictive density of the observed data. This rule can produce inferences that are incoherent, because the ‘prior’ depends on the statistical data, or absurd in other ways; see the example of a degenerate ML-2 prior in Berger (1985a, pp. 98–100).

7. Shannon and Weaver (1949), Jaynes (1957), Aczel and Daroczy (1975), Rosenkrantz (1977) and Shore and Johnson (1980).

8. For example, in measuring the amount of information provided by statistical observations concerning a possible event; see note 13 of section 5.3.

9. Classical equilibrium thermodynamics was derived in this way by Boltzmann and Gibbs; see Jaynes (1983). This use of the PME is advocated by Good (1963, 1965, Ch. 9), who applies it to the analysis of contingency tables.

10. However, the PME (or the principle of indifference) yields different hypotheses when applied to different spaces Ω which describe the same

experiment. An important example from statistical mechanics is discussed by Feller (1957, pp. 36, 39) Fine (1973, p. 169) and de Finetti (1975, section 10.3).

11. Jaynes (1968). See also Jaynes (1957) and Rosenkrantz (1977, pp. 60–1).
12. That is clear if Ω is a binary space and there is no prior information. For large n , most of the possible binary sequences yield relative frequency r_n close to $\frac{1}{2}$ for each binary outcome. However, there is no reason to expect the observed r_n to be close to $\frac{1}{2}$. Instead, we expect r_n to approximate the underlying chances. Jaynes’ view seems to be that if r_n is not close to P_0 , then Your prior information about the physical conditions of the experiment (which is used to determine \mathcal{M}) must have been incomplete. But such information usually is incomplete!
13. The proof requires some care because $\log Q$ is not a gamble. To apply the minimax theorem, restrict \mathcal{P} to a closed subset on which $P(\log Q)$ is bounded, and use a limiting argument. The logarithmic scoring rule is proper, so $P(\log Q)$ is uniquely maximized by $Q = P$.
14. The **upper entropy** of \underline{P} is defined by $\bar{H} = -\max\{P(\log Q): Q \in \mathcal{P}\} = \max\{H_p: P \in \mathcal{M}\}$. This is the minimum payment You require before You are willing to be subject to the logarithmic scoring rule, whereby You choose Q and then pay a positive penalty $-\log Q$. Similarly the **lower entropy** of \underline{P} is defined by $\underline{H} = -\max\{\bar{P}(\log Q): Q \in \mathcal{P}\} = \min\{H_p: P \in \mathcal{M}\}$. This is the maximum price You are willing to pay to subject others to the logarithmic scoring rule. Compare with the definitions of upper and lower variance in Appendix G.
15. This is a special case of the quadratic scoring rule discussed in 5.7.6, by taking $X_j = \{\omega_j\}$ and $x_j = Q(\omega_j)$. Again $P(S(Q))$ is uniquely maximized by $Q = P$. These scoring rules can be extended to a continuum of rules $S_\alpha(Q) = (\alpha + 1)Q^\alpha - \alpha \sum_{\omega \in \Omega} Q(\omega)^{\alpha+1}$, where $0 < \alpha \leq 1$. The minimax response to the rule S_α is that Q which minimizes $\sum_{\omega \in \Omega} Q(\omega)^{\alpha+1}$ over \mathcal{M} .
16. We are not advocating the minimax decision rule; any choice of Q from \mathcal{M} is reasonable here. Jaynes (1957) considers the sum of squared probabilities as an alternative to entropy.
17. That can be done most easily by applying 6.4.2. The extreme points of \mathcal{M}_1 have mass functions $(1, 0, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$ and $(\frac{1}{2}, \frac{1}{2}, 0)$. By conditioning these on $\{W, L\}$ using Bayes’ rule, we obtain $(1, 0, 0)$ and $(\frac{1}{2}, \frac{1}{2}, 0)$ as the extreme points of \mathcal{M}_2 .
18. Let $\Omega = \mathbb{Z}^+$, and let P be any probability mass function for which $P(n)$ has order $n^{-1}(\log n)^{-2}$ as $n \rightarrow \infty$.
19. The method proposed by Jaynes (1968) is to specify a group of transformations under which h should be invariant. In many problems there is no natural group of transformations. Even when there is, there may be many invariant measures, or none at all, depending on the dimensions of

the group and of Ω . See Fine (1973, Ch. 6). In the problem of an unknown chance (5.5.2), Jaynes (1968) advocates the improper Haldane density, whereas Tribus (1969, p. 261) regards the uniform density as the prior with maximum entropy.

20. Compare with our justification of coherence as a principle of consistency or rationality. There does not seem to be anything ‘inconsistent’ or ‘irrational’ about failing to maximize entropy.

Section 5.13

1. The basic ideas of the theory are clearly explained in Shafer (1976). See also the reviews by Fine (1977b) and Williams (1978). Many of the criticisms of the theory, especially concerning Dempster’s rule and the behavioural interpretation, are answered in Shafer (1981a). These issues are also discussed by Dempster (1968, 1985), Shafer (1978, 1981b, 1982a, 1982b), Williams (1978) and Walley (1987).

2. The theory loses some generality by restricting attention to lower probabilities rather than lower previsions. As seen in 2.7.3 and 4.2.3, two distinct lower previsions may agree (as belief functions) when restricted to events.

3. To prove that, verify that any belief function \underline{P} is the lower envelope of the class $\mathcal{M}(\underline{P})$ defined in note 4. In fact, all 2-monotone lower probabilities (which satisfy the complete monotonicity condition for $n = 2$) are coherent. See Walley (1981, Corollary 6.3).

4. The mathematical theory of belief functions is simpler than the general theory of coherent lower probabilities in some respects, because \underline{P} is derived from a single additive probability mass function m (although m is defined on the large space of subsets of Ω , rather than on Ω itself). All the additive probabilities in $\mathcal{M}(\underline{P})$ can be obtained by distributing the probability mass $m(B)$ over the elements of B , for each subset B . It is not necessary to adopt a sensitivity analysis interpretation of \underline{P} , by assuming that there is some ‘correct’ way of distributing the probability mass. Shafer (1979) extends the mathematical theory to infinite spaces.

5. Otherwise, identify those states ψ which map to the same set $A(\psi)$.

6. This is my interpretation of Shafer (1981a, 1982a), who uses different terminology. He regards the evidence x as an ‘encoding’, produced by some randomly chosen ‘code’ ψ , of an underlying ‘message’ that ω belongs to $A^x(\psi)$.

7. This is justified by 4.3.5, provided p_i is the posterior probability $P(\psi_i|x)$. Shafer (1981a, p. 5, note 1) also equates p_i with the prior probability $P(\psi_i)$, so that the evidence x is epistemically independent of the state ψ . See also

Shafer (1982b). This extra assumption is unnecessary. (It is also unrealistic in most problems, e.g. some states ψ may be inconsistent with x .)

8. This kind of problem is discussed by Jeffrey (1968, 1983), Russell (1948, pp. 409–13) and Shafer (1981c). It is an example of *pure* evidence, discussed by Hacking (1974, 1975) and Shafer (1978). (There is evidence to support C , but no evidence against it.)

9. As noted in 4.3.5, it is unnecessarily restrictive to assume that probabilities concerning Ψ are precise and that You have no beliefs about Ω apart from those induced by the multivalued mapping.

10. Shafer (1981a) suggests that there is a more fundamental difference in meaning between belief functions and coherent lower probabilities. That may be because he considers only a narrow (sensitivity analysis) interpretation of lower probabilities. Dempster (1968, p. 225) and Shafer (1981a, p. 15) suggest two reasons for restricting attention to belief functions, rather than coherent lower probabilities or lower previsions: only belief functions can be generated by Dempster’s multivalued mappings or can be combined by Dempster’s rule. But multivalued mappings are not the only way of constructing lower probabilities, and Dempster’s rule is not the only way of combining them.

11. The two extreme points P_1 and P_2 of $\mathcal{M}(\underline{P})$ are determined by the extra conditions $P_1(H_1 \cap H_2) = 0$ and $P_2(H_1 \cap H_2) = \frac{1}{2}$.

12. Another example is given at the end of section 4.3.5. Similar examples are given by Williams (1978, p. 380) and Walley (1987). Compare with Shafer (1981a, pp. 15–16). Belief functions are inadequate in these examples because the class of belief functions is not closed under forming lower envelopes; see Williams (1978) and Shafer (1981a, pp. 40–6). In fact, the only belief functions which satisfy constraints (a) and (b) are the additive probabilities in $\mathcal{M}(\underline{P})$, and their lower envelope \underline{P} is not a belief function. We can model two tosses of a fair coin by a belief function only if we know the precise degree of dependence between tosses.

13. More generally, if the witnesses have reliability $p = P(\psi_1|x)$ and $q = P(\phi_1|y)$, then the lower probability resulting from the combined evidence is $\underline{P}^{x,y}(C) = p + q - pq$. This formula is attributed by Shafer (1978) to J. Bernoulli and J. Lambert in the 18th century. See also Hacking (1974).

14. There has been a great deal of discussion about when Dempster’s rule is applicable. See Dempster (1967a, p. 335), Shafer (1976), the discussion of Shafer (1982a), Fine (1977b), Levi (1980, section 16.5), Williams (1978), Walley (1987), and especially Shafer (1981a, pp. 46–57). It has been claimed that the rule is applicable when the two bodies of evidence are ‘intuitively independent’, ‘unrelated’ or ‘entirely distinct’, but these notions are somewhat mysterious.

15. This condition is sufficient for Dempster’s rule to give the correct

combination $P^{x,y}$, but not quite necessary, because different pairs (ψ_i, ϕ_j) may yield the same sets $A_i \cap B_j$.

16. In many problems these judgements will not be reasonable, because the witnesses might be influenced in a similar way by circumstances other than ω , e.g. by the visibility when they thought they observed C . If You suspect that both witnesses made the same mistake then You should assess $P(\psi_2|x, y, \phi_2) > 0.5$.

17. It seems reasonable to assume that $\alpha = P(\psi_1|x, y, \phi_1, C) = P(\psi_1|x, C)$, meaning that, when C occurs, the two witnesses observe it ‘independently’. This gives precise probability $P(C|x) = P(\psi_1|x)/P(\psi_1|x, C) = 1/2\alpha$. The value $\alpha = 0.5$ yields $P(C|x) = 1$, which is absurd. To be consistent with the earlier assessments $P^x(C) = 0.5$ and $\bar{P}^x(C) = 1$, You must assess the interval $[0.5, 1]$ of values for α . This is the model considered in the next paragraph. More generally, it might be reasonable to assume that $P(\psi_i|x, y, \phi_j, \omega) = P(\psi_i|x, \omega)$, a type of independence of the two sources conditional on the states ω , but this is quite different from the conditions needed to justify Dempster’s rule.

18. It may be reasonable to assume that ψ and ϕ are independent conditional on ω , but not that they are unconditionally independent.

19. Because the factorization conditions are often complicated and difficult to check, or because the belief functions are not obtained through multivalued mappings, it is tempting to invoke a vague notion of ‘intuitive independence’ to support Dempster’s rule. That seems unwise, in view of the later examples.

20. See Dempster (1967a) and Shafer (1976) for the mathematical properties of the rule and some examples. The combination is well defined unless $\rho = 0$, in which case the bodies of evidence ‘flatly contradict’ each other.

21. This is discussed by Dempster (1967a), Shafer (1976, 1981a, 1981c) and Williams (1978). The conditional probabilities are well defined unless $\bar{P}^x(B) = 0$.

22. Williams (1978) gives an example of incoherent probabilities that are produced by Dempster’s rule.

23. This is a version of the problem discussed by Mosteller (1965, Problem 13), Jeffrey (1968), Tribus (1969, p. 77), Lindley (1971a, Ch. 3), Diaconis (1978) and Shafer (1981c). The standard ‘solution’ of the problem is to point out that precise posterior probabilities are not determined by the information given, as they depend on the unknown probability that the governor will name b when b and c are to be executed.

24. These updated probabilities agree with those produced by a naïve application of the principle of indifference. (Initially there are three possibilities for reprieve, but the governor’s answer eliminates one of these.) It is generally accepted that these probabilities are unreasonable.

25. It is difficult to imagine any other interpretation of belief functions

under which the inconsistency between prior and posterior probabilities in the preceding examples is reasonable.

26. See especially Shafer (1981a, pp. 2, 17–22, 33–5). Belief functions ‘produce degrees of belief that can be used to set betting rates without fear of Dutch book’ (p. 20). Gambles are acceptable when their lower expectations are non-negative (pp. 22, 33–5). Shafer regards the behavioural implications of belief functions as ‘only a minor aspect of their meaning’, but that is enough to support the minimal behavioural interpretation in 2.3.1.

27. See 6.4.6. Applying the GBR is equivalent to applying Bayes’ rule to all additive probabilities P in $\mathcal{M}(P)$. Applying Dempster’s rule of conditioning is equivalent to applying Bayes’ rule to all P in the smaller class $\mathcal{M}_B = \{P \in \mathcal{M}(P) : P(B) = \bar{P}(B)\}$, and this is generally incoherent because \mathcal{M}_B depends on the conditioning event B . See Levi (1980, p. 386). In example 5.13.10, both \mathcal{M}_B and \mathcal{M}_C are precise, hence so are the conditional probabilities.

Although Dempster’s rule typically gives different answers to the GBR and is therefore incoherent, it is not necessarily unreasonable in such cases. When it is reasonable to make the independence judgements that justify Dempster’s rule, these judgements provide additional information which may generate (by natural extension) conditional probabilities that are more precise than those from the GBR, and unconditional probabilities that are more precise than the initial probabilities.

28. In statistical problems, Dempster’s rule appears to be unsuitable for (a) combining beliefs based on physically independent observations, and (b) combining these with prior beliefs. These problems are investigated by Walley (1987). Dempster’s theory of statistical inference is described in (1968); see also Dempster (1966, 1967b, 1969) and Beran (1971). Shafer (1982a) develops a different approach; see also Shafer (1976, 1982b). In view of the criticisms in the discussion of Dempster (1968) and Shafer (1982a), and in Walley (1987), it does not seem possible to develop a satisfactory theory of statistical inference using Dempster’s rule.

Chapter 6

Section 6.1

1. A similar distinction has been made by many other authors, including Ramsey (1926), Pratt, Raiffa and Schlaifer (1964), Hacking (1967), Mellor (1971, p. 48), de Finetti (1972, p. 193), Teller (1973) and Shafer (1981a).

2. Under both interpretations, conditional previsions $P(X|B)$ are meaningful only when B is a well-defined event (subset of possible states). In contrast, personalist Bayesians often assert that ‘all probabilities are

conditional on Your current evidence', e.g. de Finetti (1974, section 4.1), Lindley (1971a, p. 30), von Winterfeldt and Edwards (1986, p. 96). Compare with de Finetti (1972, p. 193) and Goldstein (1985), who allow conditioning only on observable events. In order to identify Your current evidence H with an event B , You must model the process of obtaining the evidence in some detail, so that You can identify what different bodies of evidence You might have obtained. It will rarely be feasible to specify H completely, let alone to identify it with an event. Your probability assessments should be constructed from (and vary with) Your evidence H , but they are not *conditional on* H . To indicate the variation with H , it is preferable to use the notation P^H rather than $P(\cdot|H)$, or better P_t , where the subscript t refers to evidence, time, person and anything else that might influence the assessments. The Bayesian notation $P(\cdot|H)$ is especially misleading when it is coupled (as in de Finetti and Lindley) with an emphasis on Bayes' rule of conditioning as the only way of updating beliefs, because it then suggests that $P(A|H)$ can always be computed from assessments of $P(A \cap H)$ and $P(H)$.

3. De Finetti (1937, p. 108) defines conditional probability in a similar way, as a two-sided betting rate on contingent gambles. Shafer (1981a, p. 37) points out that de Finetti, in his various writings, continually conflates contingent and updated previsions. De Finetti does recognize the distinction in (1972, pp. 193–4), but he concludes that only the contingent interpretation 'is free of inextricable perplexities'. Indeed, his coherence arguments for Bayes' rule apply only to contingent previsions, and he always introduces conditional previsions by giving a contingent interpretation; see (1972, sections 5.12 and 5.31) or (1974, section 4.2). Nevertheless, de Finetti often interprets $P(X|B)$ as some kind of updated prevision, e.g. (1937, p. 119), (1972, p. 194, 210), (1974, sections 4.5.3, 4.6.1, 5.9) and (1975, section 11.2, Appendix 16), especially when he discusses statistical inference. He seems to assume (without suggesting any justification) that contingent and updated previsions should be equal. Savage (1972a, sections 2.7 and 3.5) also uses an updating interpretation.

4. Goldstein (1985, 1986) suggests that it is never the case that You observe only an event B . New evidence can rarely be identified with a subset of Your initial possibility space Ω (6.11.4), although this does seem to be possible for simple statistical experiments. (In many other problems, new evidence can be identified with an event in a possibility space that is constructed after the evidence is received (6.11.5).) Goldstein's approach is more general in that he models Your current beliefs about $P_1(X)$, Your uncertain prevision for X at a later time t_1 , without assuming that $P_1(X)$ is determined by observation of any event B . His approach is less general in that he assumes that all previsions are linear. It is certainly common that

You cannot anticipate what You might learn before t_1 , or how You would react to it. But it seems unrealistic to assume that, in these cases, Your initial beliefs concerning $P_1(X)$ are determinate; we would expect them to be highly indeterminate. Goldstein's model could be extended, as in 5.10.5, to allow imprecision in both initial and updated probabilities. See also note 11 of section 6.5.

5. You need to assess the probabilities of the observation conditional on each state ω ; see 6.11.5. When the possible observations B form a partition of Ω , these probabilities are all zero or one.
6. Updated previsions model 'dispositions to form new dispositions'. This is similar to Carnap's (1971) interpretation of credibility, which is an 'underlying permanent disposition' to form new beliefs.
7. You do not know, until after You observe B , what features of P need to be assessed most precisely. See 6.11.2.
8. The implications for updating beliefs are discussed in 6.4 and 6.11.
9. This type of extension is discussed in 6.7. See also 8.3. The second problem is discussed in 8.4.
10. If θ is unobservable, the contingent gambles cannot be settled; see 2.11.9.
11. Similar principles have been proposed by Savage (1972a, section 2.7 and D1), Pratt, Raiffa and Schlaifer (1964, axiom 5), Fine (1973, 8E) and Levi (1980, section 10.4). Such principles are rejected by Hacking (1967), Mellor (1971, Ch. 2), Shafer (1981a, 1981b) and Diaconis and Zabell (1982). The 'only if' half of the principle follows from the conglomerative principle (6.3.3).
12. Provided gambles are paid in probability currency (2.2), the time at which You receive them does not affect their value.
13. This is true only if You know in advance that You will observe either B or something inconsistent with B . (Otherwise, BZ may be non-zero even though B is not observed.) Shafer (1981a, 1981b, 1981d, 1984) argues that new evidence can rarely be anticipated in this way, and that the updating principle is unjustified when the observed event B had not been anticipated. But the updating principle applies also to this case, as follows. Suppose that after obtaining the new evidence, You model this as an event B in a partition \mathcal{B} , which consists of other evidence that could have been obtained. Let H denote the evidence You had initially, augmented by the information that an event in \mathcal{B} will be observed. If Your judgements of desirability and B -desirability are based on the evidence H , then they should be related by the updating principle. (In this case, the argument given to justify the principle shows that two hypothetical dispositions would have the same effect; the dispositions are hypothetical because You never have just the evidence H .) Your judgements of desirability (made after obtaining the new evidence, but based only on H) therefore imply judgements of B -desirability

(based only on H), which imply posterior judgements of desirability (based on the total evidence, H plus B). This ‘retrospective’ application of the updating principle supports the general updating strategy described in 6.11.5, using the generalized Bayes rule, even when B is an ‘unexpected’ observation.

For purposes of updating beliefs, the updating principle is needed only for the single event B that is actually observed. Nevertheless, it seems reasonable to extend the principle to every non-empty set B . That is because B -desirability can be displayed only in those cases where B , or something inconsistent with B , will be observed. In any such case, B -desirability of Z has the same effect as desirability of BZ .

14. Again, it is assumed that either B or something inconsistent with B will be observed.

15. The principle that updated and contingent probabilities should agree is called *de Finetti’s principle* by Shafer (1981a), although it was followed by earlier authors, e.g. Ramsey (1926, p. 192) claims that the two probabilities should ‘obviously’ be equal.

16. This information is for gamblers on the football pools.

17. This might be true even if no observation was envisaged before the later time. (Regard this as the special case of a trivial partition, $\mathcal{B} = \{\Omega\}$.) Some ways in which You might change Your probability model without obtaining new evidence were described in 4.3. Consider the example in 5.10.3 and Savage (1967), of a Bayesian who uses a uniform distribution to model his initial beliefs about ω , the tenth digit in the decimal expansion of π . If he intended to compute ω , he would not be willing to commit himself to maintain the uniform distribution. Indeed, he would not be willing to commit himself to any non-vacuous updated previsions. So the updating principle would be violated. The reason is that the initial uniform distribution is not reliable, as it is inconsistent with a more thorough analysis of the current evidence about ω . Because the Bayesian cannot yet make any reliable probability judgements about ω , both his initial and updated previsions should be vacuous; then the updating principle is satisfied.

Section 6.2

1. This includes the case of conditioning on a random variable Y , by taking \mathcal{B} to be the partition which corresponds to the possible values of Y .
2. This can be generalized to remove the constraint that $B \in \mathcal{H}(B)$, by requiring that $\underline{P}(\cdot|B)$ satisfy the general coherence condition 2.5.1 when the supremum is taken over B rather than Ω . Williams (1975a) gives a more general definition of coherence for conditional lower previsions. When all the $\underline{P}(\cdot|B)$ are linear previsions, separate coherence is equivalent to coherence axiom 3 of de Finetti (1975, Appendix 16). However, de Finetti (1974,

section 4.2) gives a different definition of coherence that allows $P(A|B)$ to take any real value when $P(B)$ is zero or unspecified. That is allowed also in Kolmogorov’s theory (6.5.8). De Finetti’s concept of coherence has been developed in different directions by Goldstein (1981, 1985), Regazzini (1983, 1987) and Williams (1975a). See also the theory of Rényi (1955).

3. Although \mathcal{H} does not necessarily contain $\mathcal{H}(B)$, the new revision $\underline{P}(\cdot|B)$ preserves all the information in the original prevision. (Whenever $X \in \mathcal{H}(B)$, $BX \in \mathcal{H}$ and $\underline{P}(BX|B) = \underline{P}(X|B)$.) Note that \mathcal{H} contains all \mathcal{B} -measurable gambles, and $Y \in \mathcal{H}$ if and only if $BY \in \mathcal{H}$ for every $B \in \mathcal{B}$. Since $\underline{P}(Y|\mathcal{B})$ is \mathcal{B} -measurable, it belongs to \mathcal{H} whenever Y does.

4. See especially axiom C11 of 6.5.3, which asserts that $G(X|\mathcal{B})$ is almost desirable.

5. If all the domains $\mathcal{H}(B)$ are linear spaces, so is the domain \mathcal{H} constructed in 6.2.4.

6. C1 may be replaced here by condition 6.2.6(h). The theorem can be extended to the case where \mathcal{H} is a convex cone, provided C1 is replaced by (C1a) if $X \geq Y + Z$ and $Z \in \mathcal{G}(\mathcal{B})$ then $\underline{P}(X|\mathcal{B}) \geq \underline{P}(Y|\mathcal{B}) + Z$. It is useful to define the gamble $\inf(X|\mathcal{B}) = \sum_{B \in \mathcal{B}} B \inf\{X(\omega) : \omega \in B\}$. With this notation C1 can be written as $\underline{P}(X|\mathcal{B}) \geq \inf(X|\mathcal{B})$, which emphasizes the similarity with P1.

Section 6.3

1. The domain \mathcal{H} constructed in 6.2.4 satisfies a stronger condition, with ‘if’ replaced by ‘if and only if’. We require only the weaker condition to allow \mathcal{H} to satisfy measurability restrictions, as in 7.3.2.
2. This is analogous to the general definition of coherence in 2.5.1, which generalizes the case where \mathcal{H} is a linear space (2.3.3). The coherence condition of Williams (1975a) is still more general (but is weaker than our condition).
3. This justification can be extended to an arbitrary partition \mathcal{B} , by imagining an experiment from which You learn only which set of \mathcal{B} contains the true state. Whatever You learned, You would be willing to accept Z . Hence You should be willing to accept Z now.
4. For examples, see 6.6.6, 6.6.7 and 6.8.5. Dempster’s rule can incur sure loss even for finite partitions; see 5.13.10 and 5.13.11.
5. This is formalized in axiom C11 of 6.5.3.
6. Using the fact that Z and BZ are equivalent after observing B , by separate coherence (6.2.4).
7. This is formalized in axiom D12 of Appendix F. Here we require the principle to hold only for a single partition \mathcal{B} , but it is extended later (in 6.8 and 7.1) to several partitions.
8. De Finetti (1972, sections 5.26–5.31, 9.5.2) and (1974, sections 4.7, 4.19,

6.9.5). He apparently accepts the updating principle, so he presumably rejects the updating version (as well as the contingent version) of the conglomerative principle. See 6.8 for further discussion of this. Another argument for the contingent version is given in 6.8.4.

9. The argument given to support the conglomerative principle can be applied directly to justify avoiding sure loss for updated previsions. If this fails, You are sure to lose by accepting $G(X) + \delta$ now and accepting $G(Y|\mathcal{B}) + \delta$ after the observation.

10. See 6.5.7 and 6.5.4.

Section 6.4

1. The GBR can be solved by numerical methods, to determine $\mu = \underline{P}(X|B)$ to any required accuracy. A simple algorithm is to let μ_0 be any estimate of μ , and define a sequence of estimates by $\mu_{n+1} = \mu_n + 2\underline{P}(B(X - \mu_n))/(\bar{P}(B) + \underline{P}(B))$. Then the sequence (μ_n) converges to μ . (The error is bounded by $c\alpha^n$, where $\alpha = (\bar{P}(B) - \underline{P}(B))/(\bar{P}(B) + \underline{P}(B)) < 1$.) Faster convergence can often be achieved by a modified algorithm based on Theorem 6.4.2. Let P_0 be any element of $\mathcal{M}(\underline{P})$. Given P_n , define $\mu_n = P_n(BX)/P_n(B)$, and let P_{n+1} be any extreme point of $\mathcal{M}(\underline{P})$ that achieves $\underline{P}(B(X - \mu_n))$. (This can be found by linear programming methods.) Then (μ_n) is a decreasing sequence that converges to μ . If $\mathcal{M}(\underline{P})$ has finitely many extreme points then P_n is eventually equal to one that achieves $\underline{P}(X|B)$, and μ_n converges to μ in finitely many steps.

2. When assumptions 6.3.1 hold and $\underline{P}(B) > 0$, the GBR is equivalent to the condition: $\underline{P}(Z|B) > 0$ if and only if $\underline{P}(BZ) > 0$. Provided $\underline{P}(Z|B)$ is interpreted as an updated prevision, this condition is a version of the updating principle for lower previsions. Thus the GBR, when regarded as an updating rule, is roughly equivalent to the updating principle. (The GBR is slightly weaker, but only because \underline{P} does not determine desirability.) Teller (1973, 1976) surveys the arguments that support updating by Bayes' rule. Shafer (1982c) outlines the history of Bayes' rule and the notation for conditional probability.

3. The result is less strange if \underline{P} is reinterpreted as a model for independent tosses of two coins, the first coin fair and the second with completely unknown bias. Let A be the event that the two outcomes are the same. Initially A has precise probability $\frac{1}{2}$. After the fair coin is tossed, A has vacuous probabilities because You are still completely ignorant about the bias of the second coin.

4. If You are completely ignorant about the two units, this is formally identical to the previous model and to the model in note 3. (Identify the first coin with the random allocation of treatments, and the second coin with the comparison between responses of the two units.)

5. For example, if the experiment was repeated for six pairs of units, and in every pair the unit given treatment 1 responded better, that would be regarded by frequentists as strong evidence that the null hypothesis (of identical treatment effects) is false. That is because the initial probability $\underline{P}(A_1 \cap A_2 \cap \dots \cap A_6 | H_0) = 2^{-6}$ is very small. After observing the allocation of treatments, the updated probability might be much greater (because You know that treatment 1 was allocated to the best units), or quite indeterminate.

6. Compare with Dempster's rule (5.13.10), which produces updated probability precisely $\frac{1}{2}$ and incurs sure loss. Similar results are obtained in the example of the indeterminate integer (5.13.11). The GBR yields $\underline{P}(A_0|B_j) = 0$ and $\bar{P}(A_0|B_j) = \frac{1}{2}$ for all j . After observing B_j , You can say only that A_0 is improbable.

7. The independence conditions that justify Dempster's rule (5.13.7) are an example of this, although they seem to be rarely applicable.

8. For example, the 'device of imaginary samples' of Good (1950, 1965, 1983) is often useful in constructing prior probabilities.

9. See 6.8.2, 6.10 and Appendix J.

10. In Example 2.7.3, let $A = \{a\}$ and $B = \{a, b\}$. Then \underline{P}_1 and \underline{P}_4 agree on events, but $\underline{P}_1(A|B) = \frac{2}{3}$ and $\underline{P}_4(A|B) = \frac{1}{2}$ when the conditional previsions are determined by the GBR. The 'three prisoners' problem (6.4.4) is an example in which $\underline{P}(\cdot|B)$ is completely determined by the lower probability \underline{P} , through the inequalities in 6.4.6. (This is because the lower probability has a unique coherent extension to all gambles.)

11. Some of these inequalities follow from the axioms of Good (1962a, p. 326). The lower bound for $\underline{P}(A|B)$ is achieved whenever \underline{P} is the natural extension of a 2-monotone lower probability (3.2.4). In that case, the updated lower probability $\underline{P}(\cdot|B)$ is also 2-monotone; see Walley (1981, Thm. 7.2).

Section 6.5

1. If $B \in \mathcal{K}$ and $\underline{P}(B) = 0$, then C6 reduces to C5 and C9 reduces to C8. (To see that, write C8 as $\bar{P}(Z) \leq \sup \bar{P}(Z|\mathcal{B})$, and apply this to $Z = X - B(\bar{P}(X|B) - \underline{P}(X|B))$, using $\bar{P}(Z|\mathcal{B}) = \bar{P}_B(X|\mathcal{B})$.) If \mathcal{B} is a continuum and $\bar{P}(Y|B)$ is a continuous function of B , then C6 reduces to C4 and C9 reduces to C7. So C6 and C9 are needed only when \mathcal{B} is a discrete partition.

2. For example, any coherent prior prevision, defined on a parameter space Θ , is coherent with any sampling model $\underline{P}(\cdot|\Theta)$. See 6.7 for details.

3. We cannot simply assume that all previsions \underline{P} and $\underline{P}(\cdot|B)$ have been separately extended to \mathcal{L} by natural extension, because this may not preserve coherence.

4. In statistical problems, for example, these axioms characterize coherence of a prior prevision \underline{P} and posterior previsions $\underline{P}(\cdot|\mathcal{X})$, all defined on the

same domain of Θ -measurable gambles. If assumptions 6.3.1 hold and \mathcal{K} contains \mathcal{H} , C10 is equivalent to any of the conditions in 2 or 3 of 6.3.5, C11 is equivalent to any condition in 5 or 7, C12 is equivalent to any condition in 10, 12 or 14, and C6 is equivalent to $\underline{P}(X - \bar{P}_B(X|\mathcal{B})) \leq 0$ for all $X \in \mathcal{H}$, or to $\underline{P}(X) \leq \bar{P}(\bar{P}_B(X|\mathcal{B}))$ for all $X \in \mathcal{H}$.

5. To see that $\underline{P}(G(X|\mathcal{B}))$ may be non-zero, let $X = H_2$ in Example 6.4.3. Then $\underline{P}(H_2|\mathcal{B}) = 0$ and $\underline{P}(G(H_2|\mathcal{B})) = \underline{P}(H_2) = \frac{1}{2}$.
6. This result holds also if \mathcal{B} is infinite but, for every gamble $X \in \mathcal{H}$, there are only finitely many $B \in \mathcal{B}$ such that X is not constant on B .
7. Separate coherence of $P(\cdot|\mathcal{B})$ is still required. (Axiom C1 plus linearity are sufficient for separate coherence.)
8. It follows that \underline{P} and $P(\cdot|\mathcal{B})$ are coherent if and only if P and $P(\cdot|\mathcal{B})$ are coherent for all $P \in \mathcal{M}(\underline{P})$. When $\underline{P}(B) > 0$, all $P \in \mathcal{M}(\underline{P})$ must determine the same $P(\cdot|B)$ through Bayes' rule.
9. This condition is used to define conglomerability with respect to a countable partition \mathcal{B} , by de Finetti (1972, section 5.26) and (1974, section 4.7).
10. For the case where $\mathcal{K} = \mathcal{H} = \mathcal{L}$, the equivalence of C14 and C15 was proved by Dubins (1975).
11. This is a special case of the theorem established by Goldstein (1983). Let $P_0(X)$ be Your prevision for X at time t_0 , and $P_1(X)$ Your prevision at some well-determined future time t_1 . (What You might learn before t_1 need not be specified as a partition \mathcal{B} .) At t_0 , You are uncertain about the value of $P_1(X)$. Suppose that You regard $P_1(X)$ as a gamble and assess a precise prevision $P_0(P_1(X))$. Then Goldstein's theorem states that $P_0(X) = P_0(P_1(X))$. This result is a simple corollary of the general conglomerative principle: if Z is sure to be desirable at t_1 , then Z is desirable at t_0 . (Apply this to $Z = X - P_1(X) + \delta$ and to $Z = P_1(X) - X + \delta$. Goldstein's assumption (vi) essentially implies this principle.) However, it is unrealistic to assume that present beliefs about the future prevision $P_1(X)$ are determinate, especially when You cannot anticipate what kind of evidence You may receive before t_1 . Allowing P_0 and P_1 to be coherent lower previsions, the general conglomerative principle implies $P_0(X - P_1(X)) \geq 0$, which generalizes axiom C11. Hence $\underline{P}_0(X) \geq \underline{P}_0(P_1(X))$. This is automatically satisfied if Your present beliefs about P_1 are vacuous, so $P_0(P_1(X)) = \inf X$.
12. To prove this, use 6.9.1 to show that C15 holds. In these two cases, given any linear prevision P , there exist linear conditional previsions $P(\cdot|\mathcal{B})$ that are coherent with P .
13. Apply C15 to AX , where $A = \{\omega : (P_1(X|\mathcal{B}) - P_2(X|\mathcal{B}))(\omega) \geq \delta\}$. Compare with Example 6.6.5, which shows that $\underline{P}(X|\mathcal{B})$ may be far from unique when linearity is not required.

14. Another difference, discussed in 6.9.8, is that we require that all previsions can be coherently extended to all gambles. That is incompatible with linearity in general.
15. For details, see Billingsley (1979, section 33) or Breiman (1968, Ch. 4). Throughout the discussion, we assume that all gambles X and Y are bounded and measurable with respect to some basic σ -field.
16. Also the coherence conditions for several partitions are stronger than C15; see 7.9. For some countably additive P and σ -fields \mathcal{A} , it is not possible to choose versions of $P(X|\mathcal{A})$ satisfying Kolmogorov's condition such that $P(\cdot|\mathcal{A})$ is a linear prevision; see Billingsley (1979, prob. 33.13).
17. Kolmogorov (1933, p. 51).
18. The meaning of contingent or updated probabilities $P(A|B)$ does not depend on the specific partition \mathcal{B} that contains B . It is the same whether $\mathcal{B} = \{B, B^c\}$ or B^c is partitioned into smaller sets. Even under a frequentist interpretation of P , it is possible to define probabilities conditional on an isolated event of probability zero. Let r_n denote relative frequency in n repetitions of an experiment, and identify probability with limiting relative frequency. Suppose $r_n(B) \rightarrow 0$, so $P(B) = 0$. Provided B occurs infinitely often, it is reasonable to define $\underline{P}(A|B) = \liminf(r_n(A \cap B)/r_n(B))$. Then $\underline{P}(\cdot|B)$ is well-defined and may be linear, but it does not depend on any choice of partition \mathcal{B} or σ -field \mathcal{A} . Nor is it determined by P .
19. As in the theories of Rényi (1955) and de Finetti (1972, 1974). It may seem that the difference between the two approaches, which is concerned with observations of probability zero, is irrelevant to real statistical problems because statistical observations are recorded in discrete form. But see sections 6.10 and 8.6 for the difficulties that arise when (as usual in statistical problems) the discrete observations are modelled as continuous variables.
20. It appears that Kolmogorov (1933, p. 47) did regard the partition \mathcal{B} , rather than the σ -field \mathcal{A} , as fundamental. However, it is now standard to admit arbitrary σ -fields.
21. Billingsley (1979, pp. 45, 381) suggests that \mathcal{A} 'can in principle be identified with an experiment or observation', by supposing that, for every $A \in \mathcal{A}$, You learn from the experiment whether or not ω is in A . But the next example, which is discussed by Billingsley (1979, pp. 388–9), shows that this interpretation is untenable. Billingsley dismisses the difficulty by suggesting that the interpretation really does not matter! Unfortunately, his attitude is typical of the Kolmogorov tradition. Compare with de Finetti (1972, 1974), who treats these issues more carefully.

Section 6.6

1. Verify that C12 holds, and C11 holds whenever $P(\cdot|\mathcal{B})$ is vacuous.

2. When $P_0(B) = 0$, conditional previsions are not determined by the GBR.
3. These exist by 3.6.8.
4. This can happen also when $P(B) > 0$ for all B , so that C12 is non-trivial, e.g. consider a convex combination of P with a countably additive distribution that is symmetric about zero.
5. This example is discussed by de Finetti (1972, section 9.5.2), who ascribes it to Dubins. The same results hold if \underline{P} is replaced by the lower prevision \underline{P} defined in 6.6.7.
6. The conglomerative principle (6.3.3) is violated for $Z = A - 0.8$.
7. This is discussed by de Finetti (1974, p. 178), who does not seem to be disturbed that Bayes' rule leads to a sure loss.
8. This is true for every coherent lower prevision \underline{P} : any mixture of \underline{P} with the vacuous lower prevision is fully conglomerable, and can therefore be coherently updated.
9. For the non-conglomerable P in 6.6.6, $\bar{P}(A|B_n) = P(A|B_n) = 1$ for all n , but $\underline{P}(A|B_n) \rightarrow 0$ (very quickly) as $n \rightarrow \infty$.
10. In fact, 6.6.4 shows that P is fully conglomerable.

Section 6.7

1. This can be done for several partitions \mathcal{B} . The most useful partitions to consider are those for which \underline{P} and $\underline{P}(X|\mathcal{B})$ can be assessed most precisely. It is often useful to introduce a parameter space Θ , because the sampling models $P(\cdot|\Theta)$ are precise.
2. See 5.3 and 5.4 for examples. You could similarly construct prior previsions concerning θ by assessing posterior previsions $\underline{P}(\cdot|x)$ for hypothetical observations x together with an \mathcal{X} -marginal \underline{P} .
3. Here the minimal extensions satisfy the conglomerative axiom C11 with equality, $E(X - \underline{E}(X|\mathcal{B})) = 0$. In the special case where \underline{P} is vacuous on \mathcal{K} , there is equality in axiom C8, $\underline{E}(X) = \inf \underline{E}(X|\mathcal{B})$.
4. One consequence is that, provided \mathcal{B} is finite or countable, any separately coherent $\underline{P}(\cdot|\mathcal{B})$ can be obtained as the unique coherent extension of some unconditional prevision \underline{E} . Let \underline{P} be any \mathcal{B} -marginal such that $\underline{P}(B) > 0$ for all $B \in \mathcal{B}$, and take $\underline{E}(X) = \underline{P}(P(X|\mathcal{B}))$ to be the minimal coherent extension of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$. Since each $\underline{P}(B) > 0$, $\underline{P}(\cdot|\mathcal{B})$ is uniquely determined by \underline{E} through the GBR. When \mathcal{B} is uncountable, $\underline{P}(\cdot|\mathcal{B})$ can be obtained from some \underline{E} as the unique coherent extension that satisfies a further regularity condition; see Appendix J4.
5. This means that a lower prevision \underline{P} on \mathcal{L} is not always determined by its \mathcal{B} -marginal and previsions conditional on \mathcal{B} . (That is true even when the \mathcal{B} -marginal is a linear prevision.) Various statistical examples are given in 7.9, 8.3 and 8.5. It may not be sufficient to specify marginal and conditional

- previsions, unless the conditional previsions are linear (as is often the case in statistical problems).
6. By properties 5 and 6 of 6.3.5.
 7. This holds for infinite \mathcal{B} , and therefore goes beyond the lower envelope theorem 8.1.10.

Section 6.8

1. If P7 holds for every partition \mathcal{B} , then P8 clearly holds. Conversely, given X and disjoint sets B_n satisfying the hypotheses of P7, replace B_1 by $B_1 \cup (\bigcap_{n=1}^{\infty} B_n^c)$ to form a partition, and apply P8 to $Y = \sum_{n=1}^{\infty} B_n X$. P8 is weaker than the similar axiom D12 in Appendix F.
2. See 6.10 and Appendix J.
3. De Finetti discusses conglomerability in (1972, sections 5.26–5.31, 9.5.2) and (1974, sections 4.19, 6.9.5). His main argument against conglomerability is that it rules out linear previsions that are not countably additive; see 6.9.7.
4. Both de Finetti and Goldstein (1983, 1985, 1986) defend non-conglomerable models by rejecting the contingent version of the conglomerative principle (6.3.3). But, assuming all previsions are linear, either the updating principle or the (original) conglomerative principle is sufficient to justify updating by Bayes' rule. Goldstein (1983) accepts the conglomerative principle, and de Finetti seems to accept the updating principle, so both seem to be committed to updating by Bayes' rule. When the initial prevision is not \mathcal{B} -conglomerable, this leads to a sure loss, and there is no other way of defining linear updated previsions that avoids sure loss.
5. For example, You can avoid sure loss by taking the updated previsions to be vacuous on B ; this satisfies the conglomerative principle but not the updating principle. Shafer (1976, 1981a) defines updated probabilities through Dempster's rule of conditioning. This violates both principles and can incur sure loss, even for finite \mathcal{B} (5.13.10).
6. The assumptions here are much stronger than those used to justify coherence (6.3.4), because we are applying the updating and conglomerative principles to every countable partition of Ω . (Alternatively, it is enough to apply the contingent version of the conglomerative principle to every countable partition.)
7. Conglomerability axioms are rejected by de Finetti (see note 3), Levi (1980, section 12.16), Hill (1980), Regazzini (1983, 1987), Hill and Lane (1986), and Kadane, Schervish and Seidenfeld (1986). These authors do not address the arguments given here, e.g. Hill (1980, p. 43) asserts that use of a non-conglomerable model 'leads to a coherent procedure in the sense of de Finetti, so there is no possibility of being made a sure loser', without recognizing that de Finetti's coherence axioms are not sufficient to rule out a 'sure loss'.

8. When \underline{P} is not specified on all gambles, we would require that it has an extension to all gambles that is coherent and fully conglomerable. This raises new problems concerning coherent extension that are not considered here.

9. This does not mean that all coherent, fully conglomerable lower previsions are reasonable models. It may be possible to justify axioms that are stronger than full conglomerability. For example, axioms of countable coherence, which extend the earlier coherence axioms to some countable sums of gambles, seem to be justified. In the case of linear previsions, these axioms are equivalent to countable additivity. So some fully conglomerable linear previsions, such as 6.6.4 and 6.6.10, may be unreasonable. (There are other problems in extending these models to linear conditional previsions; see Schervish *et al.*, 1984, Thm. 3.3.)

10. See also Hill (1980). Other examples of non-conglomerability are discussed by Stone (1976, 1982).

11. De Finetti's resolution of the problem seems to be along these lines, although it is not clear whether he favours the model in 1 or the completely different model in 4. In a practical problem there are presumably limits on the time and space available for recording the integer m , but it might be difficult to specify an upper bound M .

12. For example, P^- might be reasonable if you knew that m_2 was physically generated by a random mechanism that gave each integer zero chance of occurring. However, we know of no such random mechanism, nor of any other kind of evidence which would justify assigning upper probability zero to each state in a countable space.

Section 6.9

1. It may appear from de Finetti (1972, p. 99) that, for linear previsions, \mathcal{B} -conglomerability is equivalent to countable additivity on \mathcal{B} . But de Finetti merely shows that, if \mathcal{B} is a countable partition into sets of at least two elements, we can construct a linear prevision which is neither countably additive nor \mathcal{B} -conglomerable.

2. This follows from Schervish *et al.* (1984, Thm. 3.1). (For an alternative proof of this result, see Hill and Lane, 1986.) These authors show that the degree to which P fails to be conglomerable, measured by the supremum of $\inf \{P(A|B) : B \in \mathcal{B}\} - P(A)$ over events A and countable partitions \mathcal{B} , is equal to the degree to which P fails to be countably additive, measured by the supremum of $1 - \sum_{B \in \mathcal{B}} P(B)$ over countable partitions \mathcal{B} . They give further results which apply when P takes only finitely many values on events, but these require that conditional previsions are always linear. (For example, P in 6.6.10 is fully conglomerable but not countably additive, and P is not coherent with any linear $P(\cdot|\mathcal{B})$.)

3. This follows from 6.9.2 and the result of Ulam (1930), that (assuming the axiom of choice and continuum hypothesis) there is no countably additive probability on all subsets of an uncountable set Ω such that $P(\{\omega\}) = 0$ for all $\omega \in \Omega$.

4. Full or \mathcal{B} -conglomerability is also preserved under updating by the GBR (verify that P7 holds for the updated previsions), but not necessarily under forming convex combinations, e.g. take P^- in 6.6.6 to be zero-one valued, so both P^+ and P^- are fully conglomerable but $P = (P^+ + P^-)/2$ is not, nor under limiting operations (e.g. 6.6.7). Since one would expect convex combinations of reasonable models to be reasonable, this again suggests that some fully conglomerable models (e.g. linear previsions that are not countably additive) are not reasonable.

5. This can be done by constructing an increasing sequence of events A_n such that $P_0(A_n)$ is strictly increasing.

6. Because inner Lebesgue measure is 2-monotone, the conditional probabilities are defined whenever $\underline{P}(B) > 0$ by $\underline{P}(A|B) = \underline{P}(A \cap B) / (\underline{P}(A \cap B) + \underline{P}(A^c \cap B))$. If B is not Lebesgue measurable, $\underline{P}(\cdot|B)$ is non-additive.

7. For any such P , it is simple to construct a partition \mathcal{B} on which P is not conglomerable. Since P is not countably additive, there is a countable partition of $A = [0, \frac{1}{2}]$ (or else of $[\frac{1}{2}, 1]$) such that $\sum_{n=1}^{\infty} P(C_n) = \alpha < \frac{1}{2}$. Divide $[\frac{1}{2}, 1]$ into disjoint intervals D_n of length $P(D_n) = P(C_n)/2\alpha$. Let $B_n = C_n \cup D_n$ and $\mathcal{B} = \{B_n : n \geq 1\}$. By Bayes' rule, $P(A|B_n) = 2\alpha/(1 + 2\alpha) < \frac{1}{2}$ for all n . Since $P(A) = \frac{1}{2}$, P is not \mathcal{B} -conglomerable.

8. See especially de Finetti (1972, Ch. 5), (1974, section 3.11) and (1975, Appendix 18). Koopman (1940a), Savage (1972a, section 3.4), Dubins and Savage (1965), Levi (1980, Ch. 5), Hill (1980), Goldstein (1985, 1986), and Kadane *et al.* (1986) also reject countable additivity. Kadane *et al.* defend priors that are not countably additive on the grounds that they can reproduce inferences from improper priors, as in 7.4.9 and Heath and Sudderth (1978). As these inferences are incoherent, that is hardly a persuasive argument.

9. See 2.9.5 for example.

10. De Finetti (1972, section 5.31) is not persuaded by the 'paradoxical' consequences of non-conglomerability, because he believes that 'every new property of infinity has been considered paradoxical'.

11. Except perhaps when they take only finitely many different values on events.

12. This conclusion is supported by de Finetti's own strictures, concerning Lebesgue measure itself, against 'that which yields a unique and elegant answer even when the exact answer should instead be "any value lying between these limits".' (1974, section 6.3.3).

13. As noted earlier, rationality of P may require more than coherence and full conglomerability. The models in 6.6.9 and 6.6.10 are probably not

rational (they are not countably coherent). However, 6.9.5 and 6.9.6 are countably coherent, and the Lebesgue lower prevision certainly appears to be a sensible model.

Section 6.10

1. The reason is that $\underline{P}(BX) \leq 0$ for all gambles X . Under the correspondence between \underline{P} and desirability in 3.8, we cannot infer that any gamble contingent on B is desirable. See Williams (1975a) for a thorough discussion. However, \underline{P} contains some information about gambles contingent on B provided $\bar{P}(B) > 0$, since then $\underline{P}(BX) < 0$ for some X , and this can be used to define a non-vacuous $\underline{P}(\cdot|B)$. See Appendix J.
2. This is discussed by Kolmogorov (1933, p. 50), de Finetti (1974, section 6.9.5), Billingsley (1979, p. 391) and Hill (1980). Similar examples are given by Barnard, Jenkins and Winsten (1962), Kempthorne and Folks (1971, p. 147), Kadane *et al.* (1986), and in 8.6.5.
3. Formally, the linear prevision P is defined by $P(Z) = (4\pi)^{-1} \iint Z(\theta, \psi) \cos \theta d\theta d\psi$, for all Borel-measurable gambles Z . The coordinates θ and ψ are ‘independent’ in the sense that their joint density factorizes, but they do not necessarily satisfy our stronger criterion of independence. See 9.2.5 for discussion.
4. The discrepancy between answers 1 and 2 is ‘paradoxical’ only if one expects precise conditional probabilities to be determined by P alone. Hill (1980, p. 44) and de Finetti (1974, p. 275) suggest that the uniform density 2 is the ‘correct’ density, and even that it is reasonable to take the density of θ conditional on ψ to be uniform for every possible ψ . Since the unconditional density of θ is not uniform, that violates the conglomerative axiom and is therefore unreasonable. The uniform density could be justified by further information (as in 6.10.2) for a single value of ψ , but not for every ψ simultaneously.
5. According to Kolmogorov (1933, p. 15), ‘Infinite fields of probability occur only as idealized models of real random processes.’
6. See also 8.6.5, and note 18 of 6.5. Compare with Kolmogorov (1933, p. 51) and Billingsley (1979, p. 391).
7. Some condition of this kind is needed. Otherwise, the y -scale can be transformed for a single value of x to give inconsistent evaluations of $P(A|x)$. For instance, transform ψ in Borel’s paradox to $\alpha = \psi^3/\pi^2$ when $\theta = 0$, $\alpha = \psi$ when $\theta \neq 0$. After observing $\theta = 0$, $A = \{(\theta, \psi): |\psi| \leq \pi/2\}$ agrees with $A' = \{(\theta, \alpha): |\alpha| \leq \pi/8\}$, but the limiting argument leads to $P(A|\theta = 0) = \frac{1}{2}$ and $P(A'|\theta = 0) = \frac{1}{8}$. Because α is merely an index variable whose meaning varies with θ , it is not valid to regard $A'(\delta) = \{(\theta, \alpha): |\alpha| \leq \pi/8, |\theta| \leq \delta\}$ as a ‘neighbourhood’ of $A'(0)$.

8. This cannot hold for all gambles Z , e.g. $\underline{P}(B(x)|B(x, \delta)) = 0$ but $\underline{P}(B(x)|x) = 1$ for separate coherence.
9. To verify that $\underline{P}(\cdot|x)$ is separately coherent, use 6.4.2, 2.3.3, 6.2.7 and 2.6.7. When $\underline{P}(\cdot|x)$ is defined in this way for every possible x , further regularity conditions (such as those in 6.10.4) are needed to ensure coherence of $\underline{P}(\cdot|\mathcal{X})$ with \underline{P} .
10. It must be emphasized that the correct conditional density (if there is one) is not determined by the joint density f . There is generally a differentiable, one-to-one transformation of variables (x, y) to (x', y') such that $x' = 0$ determines the same event as $x = 0$. (By transforming coordinates in Borel’s paradox, the same great circle can be described by either $\theta = 0$ or $\psi = 0 \pmod{\pi}$.) The joint density for (x', y') can be computed from that for (x, y) by standard rules. Barnard *et al.* (1962), Kempthorne and Folks (1971, p. 147) and Kadane *et al.* (1986) give examples in which Bayes’ rule produces inconsistent answers when used to condition on the variables x and x' . But it is simply not valid to apply Bayes’ rule to an arbitrary conditioning variable. That can be justified for at most one of the two variables. Conditional probabilities are indeterminate in these examples, until further information is provided.
11. When this assumption fails, the conditional densities $g(y|x)$ given by Bayes’ rule should be multiplied by a factor proportional to $c(y) = \lim_{\delta \rightarrow 0} v(\{u: (u, y) \in B(x, \delta)\})$, which measures the imprecision of the discrete observations as a function of y . See the example in 8.6.5.
12. Subject to the condition that $P(Z|\mathcal{X})$ is measurable whenever Z is.
13. See Royden (1963), Taylor (1973) or Billingsley (1979).
14. It suffices that the function ε in 6.10.4(d) is independent of y , and $\inf\{q_y(x): y \in \Gamma\} > 0$. Then the proof of 6.10.4 extends to this case.
15. Similar results hold for intervals of measures (4.6.3, 4.6.4). Bayes’ rule can be used to update the upper and lower density functions.
16. See also the discussion in 3.8.6.
17. This is suggested by de Finetti (1972, section 5.13), (1974, section 4.18) and (1975, Appendix 18.2). See also Parikh and Parnes (1974).

Section 6.11

1. De Finetti (1974, section 4.5.3).
2. Hacking (1967), Mellor (1971, Ch. 2), Shafer (1981a, 1981b), Diaconis and Zabell (1982).
3. See 6.5.8.
4. Two examples are discussed in 6.4.
5. This is typical in cases of ‘prior–data conflict’. See Walley and Pericchi (1988) and 5.4.

6. One way to do this is to reassess the ‘prior’ after observing the data to determine the tails more precisely, and then reapply the GBR. There are other ways.
7. This is recognized by de Finetti (1972, p. 193) and Shafer (1981d, 1984).
8. The usual model is that $P(0) = P(2) = \frac{1}{4}$, $P(1) = \frac{1}{2}$, so the GBR gives $P(2|B) = \frac{1}{3}$. In the second case, it is natural to assess $P(\text{'no'}|\omega=1) = \frac{1}{2}$ and $P(\text{'no'}|\omega=2) = 1$. The more general version of the GBR (6.11.5) then yields $P(2|\text{'no'}) = \frac{1}{2}$.
9. This is emphasized by Shafer (1981b, 1981d, 1984) and Goldstein (1983, 1985, 1986). Shafer (1984), which formalizes the notion of an ‘anticipated’ observation, is especially relevant.
10. It is even more difficult to specify what other information I might have received, and to assess the probabilities of receiving exactly the information I did, conditional on the states of interest!
11. When it can be identified with a subset B , the assessments $\underline{P}(x|\omega) = 1$ if $\omega \in B$ and $\bar{P}(x|\omega) = 0$ if $\omega \in B^c$ yield the previous GBR.
12. For this case, details of the GBR and updating procedure are given in 8.4 and 8.5.
13. It may also be difficult to assess the ‘prior’ probabilities of observing x after the observation has been made.
14. The updated probabilities can be computed by the formulas in 8.5.5, e.g. if You assess $\underline{P}(A) = \frac{1}{3}$, $\bar{P}(A) = \frac{2}{3}$ and $P(x|A^c)/P(x|A) = 2$ (precisely), then $\underline{P}(A|x) = 0.2$ and $\bar{P}(A|x) = 0.5$.
15. Similarly, if You have no prior information about a horse race but learn the track prices, it is sensible to adopt these as Your upper probabilities.
16. This is discussed by Jeffrey (1983, Ch. 11) and (1968), Skyrms (1980), Shafer (1981c), Diaconis and Zabell (1982), Pearl (1988). It can be regarded as a special case of the assessment strategy discussed in 6.7, which uses assessments of \mathcal{B} -marginal previsions and previsions conditional on \mathcal{B} .
17. Typically the ‘initial’ previsions must be assessed retrospectively, after observing x . In some cases it is difficult even to model the accumulation of evidence as an ‘event’.
18. In the Bayesian case, when the initial probabilities are precise and incorrigible, the updated probabilities are determined by Bayes’ rule and there is no scope for other strategies. But if other strategies disagree with Bayes’ rule then You may want to reconsider the initial assessments. Bayes’ rule and the GBR are primarily coherence conditions rather than updating strategies; they can be used to constrain initial probabilities, as well as updated ones.

Chapter 7

Section 7.1

1. This assumption can be dropped if $G(Y_i|\mathcal{B}_i)$ are replaced by finite sums $\sum_j G(Y_{ij}|\mathcal{B}_i)$ in the following definitions.
2. When a set B belongs to two different partitions and $\underline{P}(B) = 0$, the value of $\underline{P}(Y|B)$ may differ between the partitions. In Borel’s paradox (6.10.1), the same great circle B may be regarded as a line of constant latitude or as a line of constant longitude (mod π). If we consider the two partitions into lines of constant latitude or lines of constant longitude, and define $P(\cdot|B)$ in each case by the natural limiting process (6.10.2), then $P(\cdot|B)$ will be uniform in one partition but not in the other. In such problems it is necessary to use the notation $\underline{P}_i(\cdot|B)$, where i indexes the partition in which B is embedded. That will not be necessary in the statistical problems considered here.
3. Some Bayesians seem to regard the marginal gambles $G(Y)$ and $G(Y|B)$ as acceptable because they have zero prevision. But that cannot be sufficient for acceptability, since all gambles BZ have zero prevision when B has zero probability. Non-negative gambles such as B should be acceptable, whereas non-positive (non-zero) gambles such as $-B$ should not.
4. Compare with axioms D11 and D12 of Appendix F.
5. The principle is less compelling when the partition \mathcal{B} represents the value of a statistical parameter that is not observable. In the statistical problems studied in 7.3–7.5, the ‘avoiding loss’ conditions can be supported by a different argument, concerning biased betting procedures. See 7.3 (note 4) and 7.5.4. If the conglomerative principle is rejected for infinite partitions, then the coherence condition 7.1.4(b) should be restricted to involve only finite sums of contingent gambles $G(Y|B)$. It then agrees with the coherence condition of Williams (1975a), outlined in Appendix K.
6. The gamble Z constructed in 7.1.2 might be negative only on a set of probability zero. We rule this out, for the same reasons that we require $\underline{P}(\cdot|B)$ to be separately coherent even when B has probability zero; cf the Kolmogorov approach (6.5.8). Even when $Z(\omega) < 0$ for all $\omega \in \Omega$ (a sure loss), Z may have zero prevision. Again, that does not make Z acceptable (see note 3).
7. When all the partitions are finite, uniformity is automatic. In that case, avoiding sure loss is equivalent to avoiding uniform sure loss, and avoiding uniform loss is equivalent to avoiding partial loss.
8. If $\underline{P}(Y_0|B_0)$ is interpreted as an updated prevision, the updating principle (6.1.6) is needed here.

9. Assuming that $BY \in \mathcal{K}_i$ whenever $B \in \mathcal{B}_i$ and $Y \in \mathcal{K}_i$, as in 6.3.1 and 7.3.2.
10. For example, the gambles $G(Y_i | \mathcal{B}_i)$ in 7.1.4 might be multiplied by arbitrary, non-negative, \mathcal{B}_i -measurable gambles. Stronger conditions of countable coherence or strict coherence could also be required.
11. Compare with 2.6.3 and 3.3.3.
12. This remains true if ‘coherent’ is replaced by ‘weakly coherent’. It is clear also that if $\underline{Q}(\cdot | \mathcal{B}_i)$ dominates $\underline{P}(\cdot | \mathcal{B}_i)$ for each i , and $\underline{Q}(\cdot | \mathcal{B}_1), \dots, \underline{Q}(\cdot | \mathcal{B}_m)$ avoid partial (or uniform or sure or uniform sure) loss, then so do $\underline{P}(\cdot | \mathcal{B}_1), \dots, \underline{P}(\cdot | \mathcal{B}_m)$.
13. The converse does hold when all the partitions are finite; see Appendix K.

Section 7.2

1. None of these interpretations accounts very well for the ways in which sampling models are used in practical statistics.
2. Even in these cases, the parameter is regarded as a physical quantity that is determined, in some unknown way, by properties of the experimental arrangement (e.g. by physical properties of the coin and its tossing method). It is not merely an index for the true sampling model. If the experiment is modified but the determining properties remain the same, then the sampling model may change but θ will not. For example, if θ is the chance that a coin lands heads then $P(\cdot | \theta)$ can be modified by changing the number of tosses or changing the stopping rule. Provided the tossing method remains the same, so does θ .
3. For surveys and criticisms of frequentist interpretations, see Fine (1973) and Kyburg (1970, 1974a).
4. See Fine (1973, pp. 85–9, 239), Russell (1948), Fisher (1958), and Kempthorne and Folks (1971).
5. See Reichenbach (1949), Fisher (1956a, p. 36), Salmon (1966) and Gillies (1973).
6. Different propensity interpretations are proposed by Popper (1959b, 1983), Hacking (1965), Levi (1967, 1980), Mellor (1971), Giere (1973, 1976), and in the papers by Peirce, Fetzer, Levi and Quine in Tuomela (1978). The idea dates back at least to Peirce, and Ramsey (1931, pp. 206–11) clearly distinguishes chances from finite or limiting relative frequencies. For detailed criticism of the various theories, see Kyburg (1974b).
7. These hold because chances are assumed to be precise, see 7.2.9. Hacking and (sometimes) Popper regard chances as dispositions to produce particular long-run relative frequencies in repetitions of an experiment. This view is close to the hypothetical limiting frequency one; see Popper (1983, p. 356) and Kyburg (1974b). Mellor and Levi do not tie chances so closely to relative frequencies.

8. See Mellor (1971).
9. In quantum mechanics, Schrödinger’s equation relates chances (determined by the psi function) to electrostatic potential.
10. See Kyburg (1974b) for a more thorough discussion.
11. Many authors have endorsed this principle, including Hacking (1965, p. 135), who calls it the frequency principle, Mellor (1971), Levi (1980, p. 254) and Lewis (1980), who calls it the principal principle.
12. Smith (1961) and Mellor (1971, Ch. 8) give similar arguments. Hacking (1965, Ch. 4) rejects the second kind of argument because of its reference to a ‘long run’ of trials.
13. Certainly chances depend not only on the events to which they are ascribed, but also on the kind of trial that produces the event. For example, the event that a thumbtack falls pin-up can be produced by tossing a specific thumbtack, or by tossing a thumbtack selected at random, and the event can have different chances in the two experiments. See Mellor (1971, pp. 165–6) and Levi (1980, p. 253).
14. For Mellor (1971, p. 151), propensities exist only where there is irreducible randomness: ‘If propensities are ever displayed, determinism is false.’ Suppes (1984) argues that ‘the fundamental laws of natural phenomena are essentially probabilistic rather than deterministic in character’, but he also gives many examples of complex phenomena that may be deterministic, for which stochastic models are used.
15. According to Levi (1980) and Quine, in Tuomela (1978), propensities may eventually be replaced in scientific theories by explanations in terms of the ‘micro-structure’ of experimental arrangements; they are ‘a promissory note for an eventual description in mechanical terms’ (Quine). See also Blackburn (1980).
16. For discussion, see Mellor (1971, Ch. 8) and Popper (1983, p. 28). Processes governed by simple deterministic laws can also display long-term behaviour that is extremely irregular or ‘chaotic’.
17. That is needed for practical purposes of designing buildings and establishing insurance premiums.
18. There are difficulties when chances are known relative to several subsets of the available evidence, but not relative to the total evidence (the ‘problem of the reference class’).
19. Such as maximum likelihood estimates or posterior previsions.
20. This seems to be the interpretation of Huber (1965, 1981), in his frequentist theory of robust statistics.
21. Huber (1965), Huber (1981, p. 265), and Huber and Strassen (1973) assume that the classes are disjoint. Walley and Fine (1982) show that disjointness is necessary (and also sufficient, if Θ is finite) for identifiability of θ from independent repetitions of the experiment. In the robust Bernoulli example (9.6), the classes are not disjoint and θ is not identifiable.

22. In practice, \mathcal{M}_θ might be defined as standard types of neighbourhoods, such as those in Huber (1981, p. 271).
23. This Bayesian approach to robustness is advocated by Box and Tiao (1962, 1973) and Box (1979, 1980).
24. See especially Shafer (1976), Dempster (1967a), and 5.13. This kind of indeterminacy can also result from missing or censored observations.
25. This interpretation is developed by Walley and Fine (1982, esp. section 4.3). See also Popper (1959a, sections 63–66 and Appendix iv). It is interesting that, under these frequentist interpretations, aleatory models \underline{P} must be coherent. Papamarcou (1987) has generalized von Mises' (1957) definition of a collective; this also produces coherent models \underline{P} .
26. There seems to be a dogma, accepted by frequentist statisticians, that relative frequencies in a sequence of ‘independent identical trials’ must converge. The only support for this comes from the laws of large numbers, which assume precision of the marginal probabilities. Versions of the laws of large numbers that apply to imprecise marginals do not support convergence; see Walley and Fine (1982). Note also that it is quite unnatural to adopt a sensitivity analysis interpretation of \underline{P} . Compare with the Bayesian dogmas of precision and ideal precision.
27. Mellor (1971, p. 30) suggests mathematical convenience and theoretical fruitfulness as justifications, but his account of ‘conceptual imprecision’ implies that chances are, like other physical quantities, inherently imprecise. (The laws relating chances to other physical concepts will fail to correlate perfectly, and can be reconciled only by admitting imprecision; see also Mellor, 1967.) Levi (1980, section 12.6) appeals to the principle of direct inference to justify precision of chances. Since his theory admits imprecise epistemic probabilities (albeit represented by classes of precise probability measures), it is unclear why chances should not also be imprecise. Imprecise chances would warrant imprecise epistemic probabilities through a generalized principle of direct inference. This would also support the requirement that imprecise chances (modelled by upper and lower previsions) be coherent, since epistemic probabilities should be coherent and should agree with known chances.
28. Suppes (1984, Ch. 2) gives some persuasive examples to support this view. He argues that meteorological phenomena, and perceptual phenomena such as smelling and seeing, cannot be adequately modelled by precise probabilities.
29. This is consistent with the interpretation in 7.2.7. There θ would be chosen to model the stable aspects of the process, and ψ the unstable ones. In the thumbtack example, θ might represent chances on standardized tosses, and ψ represent deviations from these due to variations in tossing method. The robust Bernoulli model in 9.6 can be interpreted in this way. It illustrates

how (and how much) we can learn about imprecise chances from repeated trials.

30. See Walley and Fine (1982).
31. General discussions and practical examples of model building can be found in Cox and Snell (1981), Mallows and Walley (1980), Leamer (1978), Box, Hunter and Hunter (1978) and Mosteller and Wallace (1964).
32. Carnap (1962, p. 574) and Ryle (1949, p. 119) appear to be instrumentalists. For detailed criticism of instrumentalism in the philosophy of science, see Nagel (1961, Ch. 6) and Popper (1962, 1983).
33. See also Dawid (1979, 1982b, 1986), section 9.5, and the discussion of interpersonal agreement in 2.11.2.
34. Because descriptive models have neither an aleatory nor an epistemic interpretation, they do not fit into the theory developed here, nor into any other formalized theory of statistics. Some examples of descriptive models and their role in statistics are discussed by Mallows and Walley (1980).

Section 7.3

1. The results extend, with minor modifications, to the case where Ω is a proper subset of $\Theta \times \mathcal{X}$. (This is needed in survey sampling, for example, where θ describes all members of a population and x describes a sample from the population.) The results can also be extended to apply to spaces larger than $\Theta \times \mathcal{X}$, e.g. $\Omega = \Theta \times \mathcal{X} \times \mathcal{Y}$, where \mathcal{Y} is the sample space for a future experiment. (We may wish to assess predictive probabilities for y after observing x .)
2. These assumptions are made to simplify the theory. They are satisfied by standard statistical models. Coherence could, of course, be investigated under more general assumptions, e.g. that $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$ are defined on different domains. (See 7.5 for one example of this.) When both Θ and \mathcal{X} are finite or countable spaces, the assumptions imply that the product σ -field contains all subsets of Ω , so that $\mathcal{F} = \mathcal{L}(\Omega)$.
3. If Ω is a proper subset of $\Theta \times \mathcal{X}$, the constraint ‘ $S(Y) = \Theta$ or $T(Y) = \mathcal{X}$ ’ should be replaced by ‘ $S(Y) \cup T(Y) = \Omega$ ’.
4. One objection to this argument is that θ is usually unobservable (2.11.9), so that gambles involving θ can never be settled with certainty. Even if such gambles are admitted, the contingent version of the conglomerative principle is needed, when Θ is infinite, to show that $Y - \frac{1}{2}\underline{P}(Y|\Theta)$ is desirable, because $\underline{P}(Y|\theta)$ must be interpreted as a contingent prevision. This principle is controversial. But there is a different argument which avoids the principle. Suppose that S1 fails for the gamble Y . Since $\bar{P}(Y|x) < 0$ for all $x \in T(Y)$, and $Y(\cdot, x) = 0$ for all other x , You will be prepared to give me the gamble Y after observing the data, whatever x is observed. But this is unfavourable

- to You, because either my lower expected gain $\underline{P}(Y|\theta)$ is positive, if the true state θ belongs to $S(Y)$, or else $Y(\theta, x) = 0$ for all possible x and I cannot lose.
5. If both Θ and \mathcal{X} are finite then S1 is equivalent to S3. If $\Omega = \Theta \times \mathcal{X}$ and either Θ or \mathcal{X} is finite then S2 is equivalent to S4.
6. These generalize axioms for linear previsions that were suggested by Stone (1976), and Heath and Sudderth (1978). Failure of S4a is called ‘strong inconsistency’ by Stone, while S4b characterizes ‘coherence’ in the sense of Heath and Sudderth. The two conditions have equivalent forms in which Θ and \mathcal{X} are reversed. The ‘consistency’ condition of Lane and Sudderth (1983) is also equivalent to S4 in the case of linear previsions. (Their condition is that there is some prior prevision P which satisfies S6* of 7.6.3.) When both Θ and \mathcal{X} are finite, S4 is equivalent to ‘expectation consistency’ as defined by Dawid and Stone (1972, 1973).
7. This is discussed by de Finetti (1972, sections 5.26–5.31), who argues that the model is a reasonable one. See 9.2.4 for further discussion. For further examples of a uniform sure loss, see 7.4.3, 7.4.4, 7.5.3 and 7.5.11.
8. Hence the conditions mentioned in note 6, which are all equivalent to avoiding uniform sure loss, are satisfied here, and are also too weak.
9. Weak coherence is equivalent to a weaker version of S5. Using the notation of 7.6.5, S5(i) is replaced by: there are no Y and θ_0 such that $\inf \underline{P}_{\theta_0}(Y|\Theta) > 0$ and $\sup \bar{P}(Y|\mathcal{X}) < 0$.

Section 7.4

- This argument is extended in 8.2.2.
- See also 8.4.7. In the Bernoulli example, $\inf P(B|\Theta) = 2^{-9}$ and the lower bound for $\underline{P}(A)$ is very small, albeit non-zero.
- Let A be the event defined in the Bernoulli example, so $A^c = (0.25, 0.75)$, and let \underline{P} be the near-ignorance prior defined in 5.3.2. Then $\underline{P}(A) > 0$ but $\underline{P}(A^c) = 0$. Indeed, $\underline{P}((\delta, 1 - \delta)) = 0$ for any positive δ . The near-ignorance model also shows that Your prior probabilities concerning future observations can be vacuous. One way of assessing non-zero values for $\underline{P}(A)$ is to use the two preceding arguments as assessment strategies. More generally, \underline{P} can be constructed from the sampling models and posterior previsions by natural extension. See 8.2 and 8.4.7 for details.
- This is shown in 8.2. For example, if some possible likelihood function is bounded away from zero, then the precise posterior determines a precise prior (8.2.7). The proper prior previsions that are implied by inferences from improper priors seem to be usually unreasonable; see the binomial and Normal examples in 8.2.
- The main reason for considering improper priors is that so-called ‘noninformative’ priors are often improper. See the discussion and references in 5.5.

- One can still interpret \underline{h} as a model for prior beliefs about Θ , either by associating it with a proper prior prevision (7.4.9) or by adopting a behavioural interpretation (5.5.4(a)). The prior beliefs seem unreasonable under both interpretations, and they are certainly not reasonable in cases of prior ignorance about Θ .
- These inferences are compatible with de Finetti’s ‘uniform’ prior on the positive integers. He allows the distributions conditional on any finite set to be uniform. Here the posterior distribution is effectively obtained by conditioning on a two-point set. Stone (1976, 1982) gives further examples of a uniform sure loss produced by a uniform density on the positive integers.
- This particular inference seems to have originated with Gauss (1809).
- Since $P(A|x) < \frac{1}{2}$ for all x , You are prepared to bet against A at even money after observing x . But this strategy is unfavourable, since $P(A|\theta) > \frac{1}{2}$ for all θ . (In a long run of repeated experiments, You are sure to lose.)
- However, these inferences are ‘almost coherent’. Example 7.8.10 shows that there are imprecise posteriors $\underline{P}(\cdot|x)$ that are arbitrarily close to the $N(x, 1)$ posteriors (uniformly in x) and coherent with the $N(\theta, 1)$ sampling models. So the $N(x, 1)$ posteriors can be regarded as (uniformly accurate) approximations to coherent posteriors. But that does not justify the $N(x, 1)$ posteriors, as even the coherent posteriors seem unreasonable (e.g. their implied prior assigns probability zero to all bounded sets). The $N(x, 1)$ posteriors also approximate the coherent posteriors in Example 7.8.3, which are more reasonable, but the approximation is not uniform in x .
- This proof is a modification of Heath and Sudderth (1978, Thm. 3), who show that the model avoids uniform sure loss.
- It might be objected that the uniform prior is inappropriate for ψ because ψ is not a location parameter (5.5.4(f)). But ψ could be a location parameter for some other observation y . If You expect to observe y and adopt uniform prior for ψ , Your prior density for θ will be very different from the uniform density You would adopt if You expected to observe x . (The two densities differ by a factor of $\frac{dy}{d\theta} = \cosh(\theta)$.) So Your prior beliefs depend strongly on what experiment You expect to observe. For a more natural example of this, see 5.5.2.
- Heath and Sudderth (1978) give weaker conditions under which the inferences avoid uniform sure loss. Their result is much more general. (It applies to generalized location families on topological groups, and the model in 7.4.6.)
- For details, see Jeffreys (1983, p. 378), Box and Tiao (1973, section 2.4), or Cox and Hinkley (1974, p. 378).
- $S2$ fails for gambles of the form $Y = B(C - \alpha)$, where B is the event that $|\bar{x}| \leq \beta s$ and α is a constant greater than γ .
- Other references concerning noninformative priors are given in 5.5.

Barnett (1982) is a good introduction to fiducial, structural and pivotal inference. References on fiducial inference include Pedersen (1978), which gives a clear account of Fisher's ideas and the connection with relevant subsets, Lindley (1958), Fraser (1961a, 1961b), Dempster (1964), Hacking (1965), Wilkinson (1977), Seidenfeld (1979), and Dawid and Stone (1982). On structural inference, see Fraser (1972, 1979). On pivotal inference, see Barnard and Sprott (1983). On right Haar measure, see Fraser (1961a, 1961b), Hartigan (1964), Nachbin (1965), Stone (1965), Hora and Buehler (1966), Villegas (1971) and Dawid, Stone and Zidek (1973). (Compare with left Haar measure, Villegas, 1977, 1981.) The models of Dempster (1966, 1968), based on multivalued mappings, also produce the incoherent inferences in 7.4.4 and 7.4.5. The inferences of Heath and Sudderth (1978) do avoid uniform sure loss in general, but that does not appear to be true of the other methods.

17. This has been suggested as a noninformative prior, but it is not universally accepted; see 5.5.2.
18. If the parameter space is reduced to the open interval $(0, 1)$, the inferences are still coherent provided the two degenerate posteriors are replaced by linear previsions concentrated in arbitrarily small neighbourhoods of 0 and 1. Lane and Sudderth (1983, Ex. 3.1) show there are no countably additive posteriors that achieve coherence in this case.
19. This is proved in Example 8.2.3.
20. Extend Y by defining $Y(0, x) = 0$ for all x . We cannot add $\theta = 1$ to the parameter space, as then the experiment would never terminate.
21. Some Bayesians, including Lindley (1965) and Box and Tiao (1973), regard the improper density as an approximation to some proper prior density such as h_n . But then the improper prior can produce misleading inferences for some values of x , e.g. when $|x|$ is large, because its posterior is a poor approximation to the posterior for the proper prior. This is discussed by Stone (1963, 1965, 1982), Hill (1974) and Dickey (1976). For illustration, see Example 7.8.3.
22. See 2.9.7. Heath and Sudderth (1978) consider such translation-invariant linear previsions as priors. They define these as weak* cluster points of the sequence $\{P_n : n \geq 1\}$, which exist because \mathcal{P} is weak*-compact.
23. Likelihood methods, which involve plotting the observed likelihood function or computing maximum likelihood estimates and observed information matrices, seem to be the best of the objective methods currently available. They are simple, easy to teach and often very revealing, especially when the sample size is large relative to the number of unknown parameters. See Fisher (1956a, Ch. 3), Barnard (1967) and Edwards (1972).
24. Other models are compared in Walley and Pericchi (1988). This approach is largely unexplored.

Section 7.5

1. The fundamental ideas are in Neyman (1937). Good accounts of the Neyman–Pearson theory are given by Neyman (1941, 1977), Cox and Hinkley (1974, Ch. 7), Mood, Graybill and Boes (1974, Ch. 8), and especially Lehmann (1986). Our discussion concentrates on Neyman's formulation of the theory, which is how the theory is usually presented. More recent formulations which emphasize conditionality, such as Cox and Hinkley (1974) and Cox (1988), seem more satisfactory.
2. See also Fisher (1956b), Wallace (1959), Yates (1964), Cornfield (1969), Robinson (1975, 1977), Bondar (1977), Pedersen (1978), Berger and Wolpert (1984) and Casella (1986). Lehmann (1986, Ch. 10) is a good summary.
3. In general, $C(x)$ is a subset of Θ but not necessarily an interval. The term 'confidence set' is more appropriate than 'confidence interval', but the latter is common usage.
4. It is difficult to prevent students from adopting this interpretation, even after it is shown to be incoherent and often absurd. According to Lehmann (1986, section 3.11), γ was interpreted as a posterior probability of coverage by eminent users of confidence intervals in the nineteenth century, including Laplace (1812), Gauss and Fourier.
5. In other words, can the 'initial precision' γ be adopted as a measure of 'final precision'? This question has been considered by many authors, including Hacking (1965), Seidenfeld (1979) and Berger (1985a).
6. A simpler, although less realistic, example can be constructed by taking Θ to be a two-point set, such as $\{0, 3\}$. The most accurate 90% confidence intervals are empty when $1.28 < x < 1.72$, and the most accurate 95% confidence intervals cover Θ when $1.35 < x < 1.65$. Other examples in which the standard confidence intervals can be empty or include all of Θ are: (a) estimating the variance ratio in the random-effects model for a one-way classification (Lehmann, 1986, section 7.11); (b) estimating a ratio of Normal means (Cox and Hinkley, 1974, Ex. 7.13); and (c) the inverse regression or calibration problem. These examples are not at all artificial – they occur in practical statistical problems. They also show that the difficulties do not arise merely from truncation of the 'natural' space Θ . The problem of empty or exhaustive confidence intervals can be resolved by reporting, instead of a single interval, a summary of the confidence intervals for all possible γ , e.g. in the form of consistency levels (7.5.13). This approach is advocated by Kempthorne and Folks (1971), Cox and Hinkley (1974), and Cox (1988).
7. This is a more extreme version of the example in Berger (1985a, p. 24). Example 7.5.11 is a less artificial version which displays the same kind of behaviour. Here θ is a location parameter, and the results are the same if \mathcal{X} and Θ are taken to be the real line. Cox and Hinkley (1974, p. 220),

Kiefer (1977) and Lehmann (1986, section 10.2) suggest that inferences about a location parameter should be conditional on the maximal ancillary statistic. That would give more sensible answers here, because B itself is ancillary, and in Example 7.5.11, but not in 7.5.2. Conditioning on ancillaries is reasonable in other problems, but it is not a general panacea. In any case, the intervals it produces are not optimal according to the Neyman–Pearson criteria, which suggests that these criteria are inappropriate.

8. This can be proved by directly applying the definitions in 7.1. (The results of 7.3 cannot be applied, because $P(\cdot|\mathcal{X})$ is defined on a much smaller domain than $P(\cdot|\Theta)$.) The theorem can be generalized to the case where $P(C|x) = \gamma(x)$ is a posterior lower probability which may depend on x (see 7.5.13), and also to the case where $P(C|\theta)$ may depend on θ but has lower bound γ .

9. Countable additivity is needed only for (ii).

10. This condition is necessary for separate coherence of $P(\cdot|\mathcal{X})$.

11. These assessments are implied by $P(C|\mathcal{X}) = \gamma$ and separate coherence of $P(\cdot|\mathcal{X})$.

12. This kind of argument is apparently due to Buehler (1959). Fraser (1977) objects that the procedure is unfair to You, because You must accept whatever gambles are chosen by Your opponent. But, according to our behavioural interpretation of probabilities, You are willing to accept the gambles involved in the procedure. If You are not willing to do so, You should not adopt γ as a posterior probability of coverage.

13. If the stronger axiom I1 fails, Your gain from the same betting procedure has negative prevision under each state θ in $J(Y)$. Also Z can be chosen so that Your gain is sure to be non-positive for all other states θ , so the betting procedure is again unfavourable.

14. This term is used by Robinson (1979a), who also gives definitions of *relevant gambles* and *semi-relevant gambles*. (Super-relevance implies relevance, which implies semi-relevance.) These are related as follows to our coherence concepts. If C incurs uniform loss then it has relevant gambles. (So incurring uniform loss is intermediate in strength between the existence of relevant and super-relevant gambles.) If there is a relevant gamble Y with $P(|Y||\theta) > 0$ for all θ , then C incurs sure loss. (Lemma 7.5.6 is a special case of this.) If the relevant gambles do not satisfy this condition, C may avoid partial loss. When C incurs sure loss, there need not be relevant gambles (e.g. 7.5.10, 7.5.11). If C incurs partial loss then there are semi-relevant gambles. If C incurs uniform sure loss then it has both positively and negatively biased *relevant selections*, in the sense of Pierce (1973), which implies that C incurs sure loss.

15. Other examples are given by Cox (1958, 1988), Berger and Wolpert (1984). For attempts to prescribe appropriate conditioning events in terms of ancillary statistics, see Fisher (1956a), Cox (1971, 1980, 1988), Cox and

Hinkley (1974), Kiefer (1977), Barndorff–Nielsen (1980), Buehler (1982) and Lehmann (1986). At present there seems to be no generally applicable theory.

16. See Fisher (1956a, pp. 35–6, 58–60, 84–6) and (1958). Fisher (1956b) criticized Welch's solution to the Behrens–Fisher problem by revealing negatively biased relevant subsets. For the solution proposed by Behrens and Fisher there are positively biased relevant subsets, hence a sure loss, but Robinson (1976) proved there are no negatively biased subsets, and Heath and Sudderth (1978) proved there is no uniform sure loss. Robinson (1982) surveys the Behrens–Fisher problem.

17. These fiducial intervals were proposed by Fisher (1935). Fisher (1956a, p. 84) denies that there can be relevant subsets here, on the grounds that the fiducial distribution is based on a sufficient statistic. See Olshen (1973) and Pedersen (1978).

18. It is uniformly most accurate unbiased or invariant.

19. See Olshen (1973) for generalizations concerning the F -test.

20. Also $P(C|B^c; \mu, \sigma) < \gamma$ for all values of μ and σ , but B^c is not a relevant subset because these probabilities are not bounded away from γ . Robinson (1976) shows there are no negatively biased relevant subsets. Heath and Sudderth (1978, Ex. 4.2) show that C avoids uniform sure loss, and this can be strengthened to show that C avoids uniform loss.

21. In the case $n = 2$ and $\gamma = 0.5$, we have $\tau = 1$, $\beta = 2.414$ and $C(x)$ is just the interval between the two observations; see Fisher (1956a, pp. 85–6). The difference between the conditional and unconditional coverage probabilities decreases as n increases, and appears to be negligible for moderately large samples. But the discrepancy for small samples is disturbing, especially as the t -distributions were developed specifically to enable satisfactory inferences to be made from small samples.

22. See Jeffreys (1983, pp. 378–83) for the Bayesian derivation. The significance of these results is discussed by Yates (1964), Cornfield (1969), Robinson (1976), Fraser (1977) and Casella (1986). There seems to be a consensus that the intervals are reasonable, even in small samples. It is especially difficult to understand how Bayesians can share this view, considering the defects of the corresponding improper prior.

23. Robinson (1979b) proves that there are no relevant gambles, and gives extensions to the case of several observations from a location family. (In that case, 'optimal' confidence intervals have relevant subsets and can incur uniform sure loss, e.g. 7.5.3 and 7.5.11.) In the case where f_0 is a symmetric density (e.g. Normal, double exponential or uniform), symmetric confidence intervals $C(x) = (x - \alpha, x + \alpha)$ apparently do avoid partial loss. This is implied by a result of Robinson (1979b); see also Lehmann (1986, p. 554).

24. Verify that $P(-B(C - \frac{1}{2}) - Z|\theta) = \exp(-\theta)/6$ if $\theta > 0$, $\exp(2\theta)/6$ if $\theta \leq 0$. Similarly there is a sure loss for $Y = B^c$ and $Z(x) = \exp(-2x)/2$ when $x \geq 0$,

- $Z(x) = 0$ otherwise. We could replace B by the event that x is less than k , for any constant k .
25. Because there are no relevant subsets for C , these rates cannot be bounded away from $\frac{1}{2}$.
26. See Pratt (1961).
27. These are uniformly most accurate unbiased or invariant; see Pratt (1961), Cox and Hinkley (1974, p. 220) and Kiefer (1977). These authors recommend different confidence intervals, based on the distribution of $\alpha + \beta$ conditional on the ancillary statistic $\beta - \alpha$, whose behaviour is much more reasonable. (As in 7.5.3, it seems essential to condition on the ancillary statistic.) When $n = 1$, the standard two-sided intervals are symmetric, $C(x) = (x - \gamma, x + \gamma)$, and these avoid partial loss by note 23.
28. There are no relevant subsets since C avoids sure loss, but $\{x_1, x_2\}$ is a relevant gamble in the sense of Robinson (1979a).
29. The discrete likelihood principle (8.6.1) and sufficiency principle are violated, because x_1 and x_2 generate the same likelihood function but lead to different inferences.
30. Pierce (1973, p. 249) conjectures that in most other problems, confidence intervals incur uniform sure loss.
31. This is advocated by Robinson (1976, 1977) and Casella (1986). It is suggested by the important example of Normal observations (7.5.8), in which there are positively biased relevant subsets but no negatively biased ones.
32. It is not clear whether the confidence intervals in Example 7.5.8 avoid sure loss or partial loss. Robinson (1976) shows there are no negatively biased relevant selections, but that is not sufficient for avoiding sure loss.
33. Neyman (1977, pp. 116–7). See also Neyman (1941, 1957). Neyman seems to envisage a single assertion or action to be based on each observation x , as if θ certainly belongs to $C(x)$. That seems inadequate for purposes of statistical inference, when no specific action is envisaged. Birnbaum (1977) argues that the Neyman–Pearson theory is usually applied for purposes of inference (summarizing the evidence concerning θ for later use) rather than decision. The ‘evidential’ interpretation of Kempthorne, Folks, Cox and Hinkley seems more suitable for inference than Neyman’s ‘behavioural’ interpretation.
34. These authors advocate reporting a nested set of confidence intervals for various levels γ . The most convenient way to summarize these is to report the observed consistency level $\text{con}(\theta, x)$ as a function of θ . (This function is comparable to the observed likelihood function, or to a Bayesian posterior density for θ .) In terms of the general relationship between confidence intervals and hypothesis tests, $\text{con}(\theta, x)$ is the smallest significance level at which the point null hypothesis θ would be rejected on observing x (sometimes called the ‘observed significance level’ or ‘ p -value’), and the

- confidence interval $C(x)$ contains just those hypotheses that would not be rejected at the $1 - \gamma$ significance level. Hypothesis testing itself relies on a consistency ordering of the possible observations, and is subject to the difficulties mentioned in the next paragraph.
35. See also Robinson (1976, 1979a), Brown (1978), Berger (1985c), Berger and Wolpert (1984) and Lehmann (1986).
36. In some problems \mathcal{B} can be defined as the set of possible values of an ancillary statistic; see the preceding references. It is generally difficult to choose \mathcal{B} so that $P(C|B; \theta)$ is independent of θ . Moreover, since the Neyman–Pearson optimality properties no longer seem appropriate, it is now unclear how the estimator C should be selected.
37. This is true whatever concept of coherence is chosen from 7.1. The non-existence of relevant gambles, in the sense of Robinson (1979a), implies that C avoids uniform loss (13), and then C is consistent (in the sense of avoiding uniform loss) with some prior prevision. Pierce (1973) gives other results concerning consistency with prior previsions.
38. This follows from Theorem 8.1.8 (note 7).
39. If $P(C|x) = \gamma(x)$ and P is a coherent linear prior, then $P(C|\Theta) - P(y|\Theta)$ has prevision zero under P (by S6*), so there is a kind of agreement between $P(C|\theta)$ and $P(y|\theta)$ in prevision.
40. For careful statements of the opposite view, see especially Cox and Hinkley (1974, sections 2.4 and 7.2), Kempthorne and Folks (1971, Ch. 13), and Neyman (1977).
- ### Section 7.6
1. Or beliefs based only on the evidence available before the experiment, if \underline{P} is constructed after the experiment (6.11.5). The case where \underline{P} is defined only on \mathcal{X} -measurable gambles is formally identical, with the roles of Θ and \mathcal{X} reversed.
 2. It was argued in 6.7 that You should be prepared to pay up to $E(Y)$ for Y . That argument is less compelling here, because θ is not observable. Provided the argument is accepted, S6 can be justified in the same way as C7.
 3. This is essentially the same as axiom C15 of 6.5.7, applied to E and $P(\cdot|\mathcal{X})$. Axiom S6* is adopted as the definition of ‘consistency’ by Lane and Sudderth (1983), and is equivalent to ‘coherence’ in the sense of Heath and Sudderth (1978). These definitions are therefore equivalent to avoiding uniform sure loss (or weak coherence), but they are strictly weaker than avoiding partial loss (or coherence) by 7.6.4.
 4. S6 does imply S3 when all posteriors are absolutely continuous with respect to the prior. More generally, if S6 holds then S3 holds for all Y such

that either $\underline{P}(S(Y)) > 0$ or $\underline{E}(T(Y)) > 0$, so that S3 can fail only on a set of prior probability zero.

5. The sure loss constructed in 7.4.4 has prior prevision zero under P , which assigns zero probability to every bounded set.

6. This is proved in 7.7.2.

7. To prove this, show that weak coherence implies S7 by appropriate choices of the gambles in 7.1.4. The three parts of S7 correspond to the three possible choices of B_0 in 7.1.4 (Ω , θ_0 or x_0 respectively). Proof of the converse is along the lines of 7.6.2. To characterize coherence, use 7.1.5 and 7.3.6.

8. The three parts of S7 can be written in various equivalent versions. For example, they are equivalent to: (i) $\underline{P}(\bar{P}(Y|\Theta)) \geq \underline{E}(\underline{P}(Y|\mathcal{X}))$, (ii) $\bar{P}(\bar{P}_{\theta_0}(Y|\Theta)) \geq \underline{E}(\underline{P}(Y|\mathcal{X}))$, and (iii) $\bar{E}(Y) \geq \underline{E}(\underline{P}_{x_0}(Y|\mathcal{X}))$, which strengthen axiom S6b. (There are versions which strengthen S6a and S6c in a similar way.)

9. Results 3 and 4 say essentially that weak coherence (S7) is unaffected if the posteriors and sampling models are modified on sets of prior probability zero. Coherence (S5) may be affected.

10. This has the interesting consequence that, for continuous models, coherence is preserved when the prior is made more precise or the sampling models and posteriors are made less precise. (Verify that S1 and S7(i) are preserved under these changes.)

11. The reason is that, when $P(\cdot|\Theta)$ is linear, \underline{E} is the unique coherent extension of \underline{P} and $P(\cdot|\Theta)$ to \mathcal{F} (6.7.3).

Section 7.7

1. The assumption here is that each probability measure $P(\cdot|\theta)$ is absolutely continuous with respect to v . Then $f(\cdot|\theta)$ is a version of the Radon–Nikodym derivative of $P(\cdot|\theta)$ with respect to v . See Royden (1963, section 11.6) or Billingsley (1979, section 32.2). Note, however, that $f(x|\theta)$ is not determined by $P(\cdot|\theta)$ and v when $v(\{x\}) = 0$, and then the posterior densities are also indeterminate. See sections 8.6.4 and 6.10.4 for conditions that justify using a particular version of $f(\cdot|\theta)$, called the *regular* density function.

2. Here q may take the value $+\infty$, but only on a set of v -measure zero since $\int q(x)v(dx) = 1$ by Fubini's theorem. Since $q(x) \leq \sup\{f(x|\theta) : \theta \in \Theta\}$, $q(x)$ is finite whenever the likelihood function is bounded. Also $q(x)$ may be zero, but not when $f(x|\theta) > 0$ for all θ .

3. The assumption that all the posteriors are absolutely continuous with respect to the prior is needed in the proof of 7.7.2 (in the last step). S1 can fail if $P(\cdot|x)$ is arbitrary when $q(x) = 0$. In some problems this assumption is unnatural, e.g. when $f(x|\theta) = 0$ unless $\theta \in A$, where $P(A) = 0$. (You may want to define $P(A|x) = 1$, whereas absolute continuity implies $P(A|x) = 0$.)

See note 5 for an example of this. In any case, because the choice of a precise posterior is arbitrary when $q(x) = 0$ or $q(x) = \infty$, it is more reasonable to adopt vacuous posteriors $\underline{P}(\cdot|x)$ in these cases. The resulting model $\underline{P}, P(\cdot|\Theta), P(\cdot|\mathcal{X})$ is still coherent. (By 8 of 7.6.8, it suffices to verify C11, C12 and S5(ii). Use the arguments in 7.7.2 and the fact that x has prior probability zero whenever its posterior is vacuous.)

4. This argument establishes more: there is no Y in \mathcal{F} such that $P(Y|\theta) > 0$ for all $\theta \in S(Y)$, $P(Y|x) \leq 0$ for all $x \in \mathcal{X}$, and $P(Y|x) < 0$ for some $x \in \mathcal{X}$. Nor can there be Y such that $P(Y|\theta) > 0$ for all $\theta \in \Theta$ and $P(Y|x) \leq 0$ for all $x \in \mathcal{X}$. (Compare with S2.)

5. Consider the limiting case $s = 0$, in which the prior density is improper. Define $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ as in the binomial model of 7.4.8. Let P be any linear prevision that assigns probability one to the event that θ is zero or one. Then S1 holds by 7.4.8, and S6* holds because both sampling models and posteriors are degenerate with probability one. So $\underline{P}, P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are coherent. All these previsions are countably additive, but this is not a standard Bayesian model as $P(\cdot|x)$ is not absolutely continuous with respect to P unless $x = 0$ or $x = n$.

6. If all singletons $\{\theta\}$ also have prior probability zero, 7 of 7.6.8 shows that virtually every posterior $\underline{P}(\cdot|\mathcal{X})$ that is dominated by the standard Bayesian posterior is coherent with P and $P(\cdot|\Theta)$.

7. Heath and Sudderth (1978) give examples involving location and scale families, including Example 7.6.4.

8. Step 3 breaks down, as $S(Y) = \Theta$, $P(Y|\theta) > 0$ for all θ , but $E(Y) = 0$.

9. It might also be argued that the evidence is often inadequate to justify a precise sampling model (7.2.7).

Section 7.8

1. If not, one option is to take $\underline{P}(\cdot|x)$ to be the vacuous posterior. Then $\underline{P}, P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ are still coherent, by note 3 of section 7.7. Alternatively, let $\underline{P}(\cdot|x)$ be the lower envelope of $P_y(\cdot|x)$ over all y such that $0 < q_y(x) < \infty$, with $\underline{P}(\cdot|x)$ vacuous if there is no such y . (Compare with the regular extension, Appendix J5.) The resulting model is not always coherent.

2. This is analogous to the beta–Bernoulli model 5.4.2. Here the degree of prior–data conflict is measured by $|x|$, the distance between the prior mean 0 and the observation x .

3. See also the robust Bernoulli model 9.6.

4. Here $\theta + v$ is identifiable from repeated observations, but θ is not. Systematic errors, such as the bias v , are often distinguished from random errors such as $x - \theta - v$.

5. Details of the GBR are given in 8.4.

6. The intervals of measures defined by DeRobertis and Hartigan (1981) are an important generalization of this model; see 4.6.4.

7. Use 6 and 8 of 7.6.8. Since $\underline{E}(\{x\}) = 0$, C12 is redundant.
8. By 3.6.3.
9. These posteriors are obtained by DeRobertis and Hartigan (1981, Ex. 4.2) from a range of improper prior measures. (This range can be regarded as a COR neighbourhood of the improper uniform prior.) These inferences are translation-invariant in the same way as the inferences from the uniform prior. Walley and Pericchi (1988) give other examples of coherent, translation-invariant inferences based on the Normal sampling model.
10. This shows that the incoherent inferences from the uniform prior (7.4.4) are almost coherent. (The posteriors $\underline{P}(\cdot|x)$ tend to the Normal $(x, 1)$ posteriors as $\tau \rightarrow 0$, uniformly in x .) Coherence (S1) is preserved when the COR posteriors are replaced by any less precise posteriors, e.g. mixtures of Normal $(x, 1)$ with the vacuous posteriors (2.9.2), or pari-mutuel neighbourhoods of Normal $(x, 1)$ (2.9.3).
11. The minimal such \underline{P} is the COR neighbourhood of the lower prevision constructed in 8.2.9. Every coherent prior prevision assigns zero upper probability to every bounded set. For that reason the posteriors appear to be unreasonable, despite their coherence.
12. See 8.4.8.

Section 7.9

1. The model avoids partial loss if and only if it satisfies S1 and the condition: $\bar{P}(G(Y|\Theta) + G(Z|\mathcal{X})) \geq 0$ for all $Y, Z \in \mathcal{F}$. (Similarly for the other types of loss: replace S1 by the appropriate axiom from 7.3.3.) However, the new condition is stronger than the two versions of C10.
2. Result 8 of 7.6.8 is a special case of this. When P is defined only on Θ -measurable gambles, as in 7.6, coherence is much stronger than pairwise coherence, even when all previsions are linear. Many posteriors will be coherent with prior beliefs about Θ (since they could be related through Bayes' rule for some sampling model), and also consistent with the sampling model (since they could be related through Bayes' rule for some prior), but the prior and sampling model together may determine a unique posterior through Bayes' rule.
3. This follows from the results of 6.5.7.
4. A third coherent extension is $\underline{E}(X) = \underline{P}(\underline{P}(X|\Theta))$, the natural extension (6.7.2).

Chapter 8

Section 8.1

1. If You wish to extend given previsions to a new prevision $\underline{E}(\cdot|\mathcal{B}_m)$, then $\mathcal{K}_m = \{0\}$ is the trivial domain, $Y_m = G(Y_m|\mathcal{B}_m) = 0$, and $S_m(Y_m)$ is the empty set. Thus the terms with $i = m$ drop out of Definition 8.1.1.

2. These generalize the properties of natural extension in the case of unconditional previsions; compare with 3.1.2.
3. It may be possible to modify the definition of natural extension to reflect these implications, but it is not clear how to do so.
4. This result (and the rest of 8.1.4) holds even if the natural extensions are not coherent. The natural extensions in Example 8.1.3 do satisfy the GBR, but they are not coherent as they violate axioms C8 and C11.
5. If \mathcal{H} contains \mathcal{K} and \mathcal{K} contains all constant gambles, this formula simplifies to $\underline{E}(X) = \sup \{\underline{P}(Y): Y \in \mathcal{K}, \underline{P}(X - Y|\mathcal{B}) \geq 0\}$ when $X \in \mathcal{H}$.
6. Formulas (a) and (b) are special cases of 8.1.4. For (c), show that $Y = 0$ and $S_2(Z)$ is either $\{B_0\}$ or empty in 8.1.1, so that $G(Y) = 0$ and $G(Z|\mathcal{B}) = G(Z|B_0)$. When $\underline{E}(B_0) = 0$, (c) defines $\underline{E}(\cdot|B_0)$ as the natural extension of $\underline{P}(\cdot|B_0)$ alone. This means that You cannot be forced to accept gambles contingent on B_0 through unconditional gambles or gambles contingent on other events. It explains why the natural extension $\underline{E}(\cdot|B_0)$ of \underline{P} is vacuous when $\underline{P}(B_0) = 0$.
7. It is clear from the proof that the theorem remains true if 'coherent' is replaced by 'weakly coherent'. See 8.2.9 for a statistical example. If the given coherent previsions $P, P(\cdot|\mathcal{B}_2), \dots, P(\cdot|\mathcal{B}_m)$ are all linear, then P has extensions to a linear prevision on \mathcal{L} that are coherent with the given previsions. In fact, these extensions are just the linear previsions that dominate the natural extension \underline{E} . (To see that any such P is coherent with the given previsions, use 7.1.4(a) for \underline{E} to verify 7.1.1 for P , and apply 7.1.5.)
8. The proof follows the lines of 8.1.8 but is more technical since we must prove coherence rather than just weak coherence. The central idea is to use the definition of natural extension to approximate each contingent gamble $B(Y - \underline{E}(Y|B))$ by a sum of gambles $\sum_{i=1}^m G(Z_i|\mathcal{B}_i)$, where $Y \in \mathcal{L}$ and $Z_i \in \mathcal{K}_i$. Since each \mathcal{B}_j is finite, any gamble $Y - \underline{E}(Y|\mathcal{B}_j)$ can thus be approximated by a finite sum of gambles $G(Z_i|\mathcal{B}_i)$. Because only finite sums are involved, coherence of the natural extensions can be verified using coherence of the specified previsions. The result also follows from Williams (1975a, Thm. 1).
- As might be expected from 8.1.8 and 8.1.9, if $P(\cdot|\mathcal{B}_1), \dots, P(\cdot|\mathcal{B}_m)$ are coherent and the partitions $\mathcal{B}_1, \dots, \mathcal{B}_k$ are finite, then $\underline{E}(\cdot|\mathcal{B}_1), \dots, \underline{E}(\cdot|\mathcal{B}_k), \underline{P}(\cdot|\mathcal{B}_{k+1}), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are coherent.
- This holds because, when all the partitions are finite, coherence agrees with Williams' (1975a) definition of W-coherence (Appendix K). In general, the extension problem is greatly simplified if only W-coherence is required. Given W-coherent previsions always have W-coherent extensions to larger domains, and these are always lower envelopes of W-coherent linear previsions. The minimal coherent extensions (if they exist) dominate the

natural extensions, which dominate the minimal W-coherent extensions. The three extensions agree if all partitions are finite. However, the minimal W-coherent extensions can be much less precise than the natural extensions, when some partition is infinite.

10. If the given previsions are all linear, then they can be coherently extended as linear previsions to \mathcal{L} , and Γ is the class of all such extensions. Compare with 3.4.2 and 3.4.3. Theorems 8.1.9 and 8.1.10 suggest that the elicitation procedure in Chapter 4 can be extended to admit assessments of conditional previsions, provided only finitely many conditioning events are involved. To develop a general theory of elicitation, it is necessary to allow arbitrary domains \mathcal{K}_i , and to require only that the assessments avoid partial loss. Their natural extension could then be constructed through a generalization of 8.1.1.

11. For example, let P be the linear prevision defined on \mathcal{L} in 6.6.10 (or the lower prevision in 6.6.9), and define $P(\cdot|\mathcal{B})$ on the trivial domain containing only the zero gamble. Then P and $P(\cdot|\mathcal{B})$ are coherent. Their natural extensions are P and $\underline{E}(\cdot|\mathcal{B})$, the vacuous conditional prevision, and these are coherent. But the natural extensions are not lower envelopes of coherent linear pairs, since there is no linear $P(\cdot|\mathcal{B})$ defined on \mathcal{L} that is coherent with P .

These examples involve linear previsions that are not countably additive. It is not known whether Theorems 8.1.9 and 8.1.10 extend to infinite partitions when the given previsions are lower envelopes of countably additive, linear previsions.

Section 8.2

1. See 6.7 and 8.1.7, identify Ω with $\Theta \times \mathcal{X}$ and \mathcal{B} with Θ .
2. See 6.8 and 8.1.6, identify \mathcal{B} with \mathcal{X} . By reversing the roles of θ and x , 1 also covers the extension of prior beliefs about x and posterior beliefs about θ to prior beliefs about (θ, x) , and 2 covers the extension of prior beliefs about (θ, x) to sampling models. The results of 6.8 also cover the extension of prior beliefs about several observables (x, y) to posterior beliefs about y after observing x . Another important type of extension, from sampling models and prior beliefs about x to prior beliefs about (θ, x) , was discussed in 2.11.9. If $\underline{P}(\cdot|\Theta)$ and an \mathcal{X} -marginal \underline{P} are defined on linear spaces \mathcal{H} and \mathcal{K} , where \mathcal{H} contains \mathcal{K} , and avoid sure loss, then their natural extension can be written (using 8.1.5) as $\underline{E}(X) = \sup\{\underline{P}(Y): Y \in \mathcal{K}, \underline{P}(X - Y|\Theta) \geq 0\}$ when $X \in \mathcal{H}$. When X is a function of θ alone, $\underline{E}(X) = \sup\{\underline{P}(Y): Y \in \mathcal{K}, \bar{P}(Y|\Theta) \leq X\}$.
3. For purposes of constructing beliefs about θ , it is not necessary that the

contemplated statistical experiment can be carried out. It may be merely a thought-experiment, as suggested by Good (1950, 1962b, 1965, 1976). You should consider experiments for which the sampling models and posterior previsions can be assessed most precisely.

4. A necessary and sufficient condition for \underline{E} to be vacuous on \mathcal{K} is that $\sup P(Y|\Theta) \leq \sup \bar{P}(Y|\mathcal{X})$ for all Y in \mathcal{F} .
5. Another possibility is to adopt vacuous posteriors when $x = 0$ or $x = n$. The following argument can be modified to give the same conclusions, that the natural extension assigns zero upper probability to the events $1 \leq x \leq n - 1$ and $A(\delta)$.
6. See Lane and Sudderth (1983, Ex. 3.1). If P is not countably additive, it may be concentrated in arbitrarily small neighbourhoods of $\theta = 0$ and $\theta = 1$, without being degenerate at these points.
7. Since $P(x = 0) > 0$ and $P(A(\delta)) = 0$ imply $P(A(\delta)|x = 0) = 0$ for all positive δ , and similarly when $x = n$.
8. This is an example of the sort of reasoning suggested in 1.7, where probability assessments are criticized by examining their implications for other gambles, through natural extension. See also 8.2.9.
9. Since $\underline{P}(\{x\}) > 0$ by coherence, \underline{P} determines a unique posterior through the GBR. If \underline{P} is linear, so is the posterior.
10. This is not true when \mathcal{X} is uncountable; then linear priors can be coherent with vacuous posteriors.
11. Similar results in Heath and Sudderth (1978) and Lane and Sudderth (1983) concern weak coherence. In view of 7.1.5, these can be proved in essentially the same way.
12. For example, Freedman and Purves (1969) show that, when both Θ and \mathcal{X} are finite, for $P(\cdot|\Theta)$ and $P(\cdot|\mathcal{X})$ to avoid sure loss there must be a linear prior P such that $P(\cdot|x)$ can be obtained from P and $P(\cdot|\Theta)$ through Bayes' rule.
13. For example, consider 8.2.3 with Θ reduced to the open interval $(0, 1)$. See Lane and Sudderth (1983), who give regularity conditions under which there is a countably additive prior.
14. Freedman and Purves (1969) establish uniqueness under stronger assumptions. There appears to be uniqueness in many statistical problems, e.g. when the sampling models have densities $f(x|\theta) = L_x(\theta)$, with respect to a σ -finite measure v , that are everywhere positive and continuous in (θ, x) . Assuming that P is countably additive, it is uniquely determined on compact sets A (in \mathcal{K}) by $P(A) = P(AL_x^{-1}|x)/P(L_x^{-1}|x)$, for every x outside a set of v -measure zero. Non-uniqueness seems to occur only when $f(x|\theta) = 0$ for many values of x and θ .
15. It also holds for $A = [0, 1 - \delta]$ or $A = [\delta, 1]$, by taking $x = 0$ or $x = n$.
16. See note 7 of section 8.1. As shown in 7.8.10, coherence can be achieved

by forming constant odds-ratio neighbourhoods of the Normal posteriors. The natural extension of this model, \underline{E} , is then coherent with $P(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{X})$. In fact, \underline{E} is just a constant odds-ratio neighbourhood of \underline{E} . Coherence of the model is preserved when \underline{E} is replaced by any dominating prior (such as \underline{E}), or when $\underline{P}(\cdot|\mathcal{X})$ are replaced by any less precise posteriors. The prior \underline{E} is translation-invariant, but not all its dominating linear previsions are, and none of these linear previsions is countably additive.

17. This is a kind of lower limit of uniform distributions on finite intervals $(u - c, u + c)$. It can be verified that the limit exists, by partitioning long intervals of length $2c$ into many shorter intervals.

18. Essentially, this follows from 8.2.5.

19. Let $E_{u,n}$ denote expectation under the uniform distribution on $(u - n, u + n)$. Then P belongs to $\mathcal{M}(E)$ if and only if, for every measurable X , there is a sequence of reals u_n such that $E_{u_n,n}(X) \rightarrow P(X)$ as $n \rightarrow \infty$.

Section 8.3

1. This also covers the case where prior beliefs are assessed concerning x rather than θ , by reversing the roles of θ and x .
2. The proof is similar to 8.2.1.
3. When X is in \mathcal{X} , this gives $\underline{E}(X) = P(X)$ since $\underline{P}(X|\Theta) = \bar{P}(X|\Theta) = X$. This confirms that \underline{E} is an extension of \underline{P} .
4. Use property 4 of 6.3.5.

Section 8.4

1. According to de Finetti (1975, p. 199), ‘The vital element in the inductive process, and the key to every constructive activity of the human mind, is then seen to be Bayes’ theorem.’ See also de Finetti (1972, p. 194).
2. It is not always useful. See 6.11 for discussion.
3. The case $\underline{P}(\{x\}) > 0$ follows from 8.1.4. For the case $\underline{P}(\{x\}) = 0$, use Definition 8.1.1 with $\mathcal{B}_1 = \{\Omega\}$, $\mathcal{B}_2 = \Theta$, $\mathcal{B}_3 = \mathcal{X}$ and $B_0 = \{x\}$. Use $\underline{P}(\{x\}) = 0$ and separate coherence of $\underline{P}(\cdot|\Theta)$ to show that $Y_1 = Y_3 = 0$ and $Y(\theta, x) > \alpha$ for all θ . Hence $\underline{E}(Y|x) \leq \inf Y(\cdot, x)$. The reverse inequality holds by 8.1.2(a).
4. To prove this, let $B = A \cap \{y\}^c$, show $\bar{P}(BY) \leq \sup \bar{E}(BY|\mathcal{X}) \leq 0$ and $\bar{P}(\{y\}Y) \leq \underline{P}(\{y\})\bar{E}(Y|y) < 0$, hence $\bar{P}(AY) \leq \bar{P}(\{y\}Y) + \bar{P}(BY) < 0$.
5. But introducing $\underline{P}(\cdot|\Theta)$ can rule out some posterior previsions that are coherent with \underline{P} alone. (The natural extension satisfies S5, but other posteriors may not.) As an example, consider 7.6.4.
6. This follows from 8.1.4 when $\underline{E}(\{x\}) > 0$, and from 8.4.1 otherwise.

7. Of course $\underline{E}(\cdot|\mathcal{X})$ is vacuous in this case.
8. By 6.9.4, it is enough that the \mathcal{X} -marginals are countably additive.
9. It suffices to define $\underline{E}(Y|x)$ when Y is in \mathcal{X} , since it is then determined for all Y in \mathcal{F} by $\underline{E}(Y|x) = \underline{E}(Y(\cdot, x)|x)$.
10. When the sampling models have density functions $f(\cdot|\theta)$ with respect to a common measure, the likelihood function is usually defined by $L_x(\theta) = f(x|\theta)$. With this definition, posterior previsions can be defined through the last form of the generalized Bayes rule. These posteriors will often be reasonable and non-vacuous, but it is necessary to check their coherence with P and $P(\cdot|\Theta)$. See 7.8.6–7.8.11 for discussion and examples. Unbounded functions L_x are not gambles and can cause trouble. (See note 12 of section 8.6.)
11. The GBR can be solved by modifications of the numerical algorithms in section 6.4 (note 1). In the first algorithm, define $\beta_{n+1} = \beta_n + 2\underline{P}((Y - \beta_n)L_x)/(\bar{P}(L_x) + \underline{P}(L_x))$. In the second, define $\beta_n = P_n(YL_x)/P_n(L_x)$ and find an extreme point P_{n+1} that achieves $\underline{P}((Y - \beta_n)L_x)$.
12. Again, $\underline{E}(\cdot|\mathcal{X})$ is not automatically coherent with P and $P(\cdot|\Theta)$. Coherence is equivalent to \mathcal{X} -conglomerability of E , defined by $E(Y) = P(P(Y|\Theta))$. The posteriors are linear whenever $P(L_x)$ is positive. Otherwise, \mathcal{X} -conglomerability of E is necessary but not sufficient for there to be coherent linear posteriors. By 6.9.1, it is sufficient that \mathcal{X} is countable and E is countably additive on \mathcal{X} .
13. As in 6.5.7, linear posteriors that are coherent with P and $P(\cdot|\Theta)$ are almost unique. When P and $P(\cdot|\Theta)$ are countably additive, $P(Y|\mathcal{X})$ is determined up to a subset of \mathcal{X} with prior probability zero. Assuming also that $P(\cdot|\theta)$ have densities $f(\cdot|\theta)$ with respect to a common σ -finite measure, S6* and Fubini’s theorem can be used to show that a linear posterior is determined by Bayes’ rule $P(Y|x) = P(YL_x)/P(L_x)$ for each Y in \mathcal{X} and all x outside a set of prior probability zero, where $L_x(\theta) = f(x|\theta)$.
14. That is true even if the likelihood function L_x is positive for only one value of θ . Compare with the regular posteriors for the vacuous prior, defined in Appendix J5, which are then degenerate at this value of θ .
15. Similarly, $\bar{E}(A|x) = 0$ when $\bar{P}(A) = 0$ and $\underline{P}(L_x) > 0$.
16. This supports a further rationality requirement of *strict coherence*, that the prior probabilities $\underline{P}(A)$ are positive for all open subsets of Θ . When Θ is countable and has the discrete topology, this means that every possible θ has positive prior lower probability.
17. This follows from 6.4.2, or from the generalized Bayes rule in 8.4.4 by writing \underline{P} as the lower envelope of $\mathcal{M}(\underline{P})$. Although $\underline{E}(\cdot|x)$ can always be written as a lower envelope of linear posteriors, a coherent model \underline{P} , $P(\cdot|\Theta)$, $\underline{E}(\cdot|\mathcal{X})$ cannot always be written as a lower envelope of coherent linear

models P , $P(\cdot|\Theta)$, $P(\cdot|\mathcal{X})$. (It can if both Θ and \mathcal{X} are finite, by 8.1.10.) Consider Example 6.6.10, where (m, n) is identified with (x, θ) , P is linear and $E(\cdot|\mathcal{X})$ is vacuous.

18. But they are the only coherent posteriors that satisfy the regularity axiom C16. See Appendix J6.

Section 8.5

1. To prove the last assertion, note that the posteriors $P(\cdot|\mathcal{X})$ coherent with Q and $P(\cdot|\Theta)$ are minimized (using 8.4.2) by defining $P(\cdot|\mathcal{X})$ from Q by 8.4.1, and then by minimizing Q . The existence of a minimal Q follows from 6.9.3 and 7.1.6.

2. This $E(\cdot|\mathcal{X})$ is also the natural extension of P and $P(\cdot|\Theta)$. That follows from 8.1.4 when $E(\{x\}) > 0$, and otherwise in the same way as 8.4.1.

3. I do not have examples where E is not \mathcal{X} -conglomerable, but some other extension Q of P and $P(\cdot|\Theta)$ is \mathcal{X} -conglomerable. Example 8.1.3 indicates how this might happen. If it does, the minimal coherent posterior is characterized by 8.5.1 and may differ from the natural extension $E(\cdot|\mathcal{X})$.

4. For examples, see 7.8.5 (continued in 7.9.3, 8.3.4), and 8.5.5.

5. To obtain these, apply Theorem 6.4.6 to E and \bar{E} , with $B = \{x\}$. The posterior probability $E(A|x)$ is equal to the lower bound $P(AL)/(P(AL) + \bar{P}(A^cU))$ whenever the prior prevision P is the natural extension of a 2-monotone lower probability (3.2.4). In that case, the posterior lower probability $E(\cdot|x)$ is also 2-monotone. See Walley (1981, Thm. 8.3).

6. The *regular extension* of P and $P(\cdot|\Theta)$ is defined by the formulas

$$\begin{aligned} R(Y|x) &= \max\{\beta: E(\{x\}(Y - \beta)) \geq 0\} = \max\{\beta: P((Y - \beta)Z(Y, \beta)) \geq 0\} \\ &= \inf\{P(YZ)/P(Z): P \in \mathcal{M}(P), L_x \leq Z \leq U_x, P(Z) > 0\}, \end{aligned}$$

provided $\bar{P}(U_x) > 0$. This may be non-vacuous, unlike the natural extension, when $P(L_x) = 0$ but $\bar{P}(U_x) > 0$. Again, the regular extension depends on the sampling model only through the upper and lower likelihood functions. See Appendix J.

7. This does not mean that coherent P , $P(\cdot|\Theta)$, $E(\cdot|\mathcal{X})$ are lower envelopes of coherent linear previsions P , $P(\cdot|\Theta)$, $P(\cdot|\mathcal{X})$. For example, 6.6.9 can be reinterpreted as a statistical model by setting $\Theta = \{+, -\}$, $\mathcal{X} = \mathbb{Z}^+$, and identifying $(+, n)$ with n , $(-, n)$ with $-n$. In this case $P(L_x) > 0$ for all x , but there are no dominating linear previsions P , $P(\cdot|\Theta)$, $P(\cdot|\mathcal{X})$ that are coherent. The lower envelope result does hold when both Θ and \mathcal{X} are finite, by 8.1.10.

8. Here the posterior is generated by natural extension. Other coherent posteriors may have $P(A|x)$ precise when the likelihood function is imprecise, as in Examples 8.3.4 and 8.5.5.

9. Whether $E(\cdot|\mathcal{X})$ is the unique coherent extension depends essentially (for discrete \mathcal{X}) on whether P has a unique extension to a joint prevision for (θ, x) that is coherent with $P(\cdot|\Theta)$. The joint prevision may be unique even if $P(\cdot|\Theta)$ is non-linear. In this example, and in 8.3.4, it is not unique.

Section 8.6

1. There is an extensive literature concerning the likelihood principle. See especially Birnbaum (1962, 1969), Cox and Hinkley (1974, Ch. 2), Basu (1975), Berger (1985b), Berger and Wolpert (1984), and other references therein.

2. They should also lead to the same decisions, provided θ includes all unknowns that are relevant to the decision.

3. In this example, frequentist inferences (such as significance tests or confidence intervals) would depend on which experiment generated the data, thereby violating the likelihood principle. The same is true of Bayesian inferences based on noninformative (or ‘reference’) priors, where these depend on the sampling model; see 5.5.2, 5.5.4(f) and 7.4.8.

As an example of how frequentist inferences depend on the experiment, suppose that You wish to test the null hypothesis H_0 that $\theta \geq 0.5$, against the alternative that $\theta < 0.5$, and You obtain the data $m = 1$ and $n = 6$. Let p denote the observed significance level for testing H_0 . If the first experiment was performed, a frequentist calculates $p = \frac{7}{64} = 0.11$, the probability when $\theta = 0.5$ of observing no more than one success in 6 trials. For the second experiment, he calculates $p = \frac{1}{32} = 0.031$, the probability when $\theta = 0.5$ that it requires at least 6 trials to obtain a success. The data would be regarded as strong evidence against H_0 if obtained from the second experiment (H_0 can be rejected at the 0.05 level), but not if obtained from the first. (A similar example is discussed by Lindley and Phillips, 1976.) Similarly, the reference priors of Bernardo (1979), Box and Tiao (1973) yield posterior probabilities $P(H_0|x) = 0.047$ for the first experiment, but $P(H_0|x) = 0.022$ for the second experiment. For comparison, the uniform prior for θ yields $P(H_0|x) = 0.0625$, and the near-ignorance prior (5.3.2) with $s = 2$ yields $P(H_0|x) = 0.0078$ and $P(H_0|x) = 0.227$.

4. Previsions may not be adequate models for some purposes, as suggested in 3.8.6.

5. By Theorem 8.2.1.

6. The prior P should obviously depend on the prior information about Θ . Ideally, P would be uniquely determined through a thorough analysis of the prior information. In practice, the extent of this analysis might depend on the data, as suggested in 5.4.5 and 6.11.2, and then P might depend on

the data. However, it does not seem reasonable for P to depend on any features of the experiment except for the observed likelihood function, so assumption 4 remains valid. The noninformative priors proposed by Box and Tiao (1973), Zellner (1977) and Bernardo (1979) do depend on other features of the experiment; see 5.5.2.

7. This is a requirement of strict coherence.

8. The justification does not apply when the likelihood function is zero on the support of the prior, so that assumption 5 fails, although I know of no reason to violate the principle in that case. (One would probably reassess the prior.) There are other persuasive arguments in favour of the discrete likelihood principle, notably Birnbaum's (1962) demonstration that it is implied by the conditionality and sufficiency principles. See Basu (1975), Berger (1985b), and Berger and Wolpert (1984) for other arguments.

9. Berger and Wolpert (1984, section 3.4) defend a continuous version of the likelihood principle, but this has the same defect. (It may be violated on observing x which has probability zero under each sampling model.)

10. For further discussion, see especially Kempthorne (1966), Barnard (1967) and Kempthorne and Folks (1971, section 10.5.2). The problems arising from the introduction of continuous models, and especially the definition of an appropriate likelihood function, are very important, but they seem to be ignored in most of the statistical literature.

11. If x is rounded to m decimal places then $\delta = 10^{-m}/2$. There is no difficulty in allowing δ to depend on x .

12. The requirement of *uniform* convergence is needed to ensure that the posterior previsions generated by the discrete likelihood functions converge to those generated by the continuous likelihood function, whatever the prior predition. It effectively rules out cases where the continuous likelihood function is unbounded.

13. These conditions are closely related to the ones in 6.10.4, which justify Bayes' rule for computing posterior densities. The same conditions are sufficient when δ is a function of x . If conditions 1 and 2 hold, but $\delta(\theta, x)$ is a known function of θ and x , then the appropriate continuous likelihood function is $L_x(\theta) \propto v(B(x, \delta(\theta, x)))f(x|\theta)$. When $B(x, \delta) = (x - \delta, x + \delta)$ this gives $L_x(\theta) \propto \delta(\theta, x)f(x|\theta)$.

14. Luis Pericchi has pointed out that this example is slightly irregular because the support of $f(\cdot|\theta)$ depends on θ . The same indeterminacy occurs in more regular problems. For example, let the sampling models be Normal $(0, \theta^2)$, and consider the two models for imprecise measurement in 8.6.6 and 8.6.7. Alternatively, interpret Borel's paradox (6.10.1) as a statistical model, by regarding θ as the parameter and ψ as the observation, and consider the models in 6.10.2. The choice of Θ is also inessential to the uniform example; the conclusions are unchanged if $(0, 1)$ is replaced by a two-point space such as $\{0.5, 1\}$.

15. If there are several decimals of equal length, report the smallest.
16. We are not suggesting that the posterior distribution $P(\cdot|x)$ may be uniform on $(x, 1)$ for all x in $(0, 1)$. That is incoherent, because when $A = \{(\theta, x): x < 2\theta - 1\}$, $P(A|\theta) = 2 - \theta^{-1}$ if $\theta > 0.5$, $P(A|\theta) = 0$ otherwise, giving $P(A) = \int_0^1 P(A|\theta) d\theta = 1 - \log 2 = 0.307$, whereas $P(A|x) = 0.5$ for all x . (C14 and C15 are violated.) But You could (coherently) adopt the uniform posteriors for all terminating decimals x , i.e. for all possible *reported* observations, as these form a set of prior probability zero.
17. Although the regular density functions are limits of discrete probabilities, as in (2) of section 8.6.4, they are the 'wrong' densities for defining a likelihood function or computing a posterior density through Bayes' rule. From condition 2 of 8.6.4, the 'right' densities $f(x|\theta)$ can be defined as limits of $P(C(\delta)|\theta)/v^2(C(\delta))$ as $\delta \rightarrow 0$, where $C(\delta)$ is the discrete event that was actually observed (regarded as a subset of $\Theta \times \mathcal{X}$), and v^2 is a measure on subsets of $\Theta \times \mathcal{X}$ (such as two-dimensional Lebesgue measure). Because $v^2(C(\delta))$ does not depend on θ , this leads to the correct likelihood function.
18. This model is examined in more detail by Kempthorne (1966, appendix B), who compares the likelihood functions for discrete and continuous observations.
19. Fraser (1968, 1972) and Barnard (1980) suggest that further 'structural' information about the experiment is needed.
20. For x_1 and x_2 to generate the same posterior by natural extension whatever prior is chosen, it is necessary and sufficient that they generate proportional upper and lower likelihood functions, provided Θ has at least 3 elements. When Θ has 2 elements, it is necessary and sufficient that x_1 and x_2 generate the same upper and lower likelihood ratios (section 8.5.5).

Chapter 9

Section 9.1

1. Compare with Keynes (1921, pp. 59, 150, Ch. 16). Independence is a symmetric relation; if A and B are independent then so are B and A , A and B^c , etc. Irrelevance is not symmetric. In fact, it is possible for seven of the eight equalities in 9.1.1 to hold but the eighth to fail, even when all events have positive lower probability. Weaker judgements of non-negative relevance and dependence are defined in 9.7.2.
2. Compare with the definition of independent experiments (9.2.1).
3. De Finetti (1974, sections 4.9, 6.9) seems to regard independence only as a property of pre-specified previsions. Independence judgements are often made prior to, or separately from, any quantitative assessments of previsions.
4. This can happen also when the probabilities are non-vacuous. If You know only that the chances of the events $A \cap B$, $A \cap B^c$, $A^c \cap B$ and $A^c \cap B^c$

all lie between 0.2 and 0.3, then You obtain the model in 9.1.6, under which A and B are epistemically independent. But they need not be physically independent; their degree of physical dependence $P(A \cap B)/P(A)P(B) - 1$ can be anywhere between -0.2 and $+0.2$.

Boole (1854) advocates a much stronger principle, that You should regard two logically independent events as epistemically independent when You know their chances but know nothing about the physical dependence between them. Boole's principle seems unreasonable. (Consider Example 6.4.3.)

5. See Fine (1973, pp. 80–2) for further discussion.

6. If A and B are both essentially trivial then they are epistemically independent, even if they are not logically independent. We might want to strengthen 9.1.1 by requiring also that A and B are logically independent. De Finetti (1974, section 4.11) strengthens the standard definition of independence by requiring logical independence, but this does not go far enough to avoid the difficulties noted in 9.1.4 and 9.2.

7. Example 9.1.6 shows that all these inequalities may be strict. The inequalities imply the 2-monotonicity property for independent events, $\underline{P}(A) + \underline{P}(B) - \underline{P}(A \cup B) \leq \underline{P}(A)\underline{P}(B) \leq \underline{P}(A \cap B)$.

8. Generally \mathcal{M} will not be convex, because independence is not preserved under convex combinations. Whenever both A and B have imprecise probabilities, $\mathcal{M}(\underline{P})$ contains some linear previsions under which A and B are not independent. Then \mathcal{M} must be smaller than $\mathcal{M}(\underline{P})$.

9. If A and B are independent under \underline{P} then they are independent under some extreme points of $\mathcal{M}(\underline{P})$. (In fact, they are independent under all linear previsions which achieve the upper or lower probability of A or B . Every P in $\mathcal{M}(\underline{P})$ satisfies $P(A) = P(B)P(A|B) + (1 - P(B))P(A|B^c)$, $P(A|B) \geq \underline{P}(A|B) = \underline{P}(A)$ and $P(A|B^c) \geq \underline{P}(A|B^c) = \underline{P}(A)$. To achieve $\underline{P}(A)$, P must satisfy $P(A|B) = P(A|B^c) = P(A) = \underline{P}(A)$, so that A and B are independent under P .) Thus, in Example 9.1.6, A and B are independent under the four extreme points (permutations of $(0.3, 0.3, 0.2, 0.2)$) which achieve $\underline{P}(A)$, $\bar{P}(A)$, $\underline{P}(B)$ or $\bar{P}(B)$. They are not independent under the two other extreme points.

10. Writing $P(\{j\}) = p_j$, the independence constraints imply that some $p_j = 1$, or else $p_2 = p_3 = p_4 = p_5$ and hence $p_1 = 1$, which is incompatible with $p_j \leq \frac{1}{2}$.

Section 9.2

1. This clearly depends on the choice of σ -fields \mathcal{C} and \mathcal{S} . The σ -fields must contain all subsets of \mathcal{X}_1 and \mathcal{X}_2 when these spaces are finite or countable, by assumptions 7.3.2, and then $\mathcal{F} = \mathcal{L}(\Omega)$. Independence of two partitions of Ω , \mathcal{B}_1 and \mathcal{B}_2 , can be defined in a similar way, by requiring $\underline{P}(Y|\mathcal{B}_2) = \underline{P}(Y)$

and $\underline{P}(Z|\mathcal{B}_1) = \underline{P}(Z)$ for all \mathcal{B}_1 -measurable Y and \mathcal{B}_2 -measurable Z . This includes the case of independent events (9.1.1) by taking $\mathcal{B}_1 = \{A, A^c\}$ and $\mathcal{B}_2 = \{B, B^c\}$. Two gambles X_1 and X_2 would be called independent when the partitions they generate are independent.

2. When $Y \geq 0$, Y is \mathcal{C} -measurable and Z is \mathcal{S} -measurable, independence of the marginals implies that $\underline{P}(YG(Z)) = \underline{P}(G(YZ|\mathcal{X}_1)) \geq 0$. Hence

$$\underline{P}_1(Y\underline{P}_2(Z)) \leq \underline{P}(YZ) \leq \bar{P}_1(Y\bar{P}_2(Z))$$

and

$$\bar{P}_1(Y\underline{P}_2(Z)) \leq \bar{P}(YZ) \leq \bar{P}_1(Y\bar{P}_2(Z)).$$

For events $A \in \mathcal{C}$ and $B \in \mathcal{S}$, we obtain

$$\underline{P}_1(A)\underline{P}_2(B) \leq \underline{P}(A \times B) \leq \bar{P}_1(A)\bar{P}_2(B) \leq \bar{P}(A \times B) \leq \bar{P}_1(A)\bar{P}_2(B).$$

Walley and Fine (1982, section 4.3) suggested the independence conditions $\underline{P}(A \times B) = \underline{P}_1(A)\underline{P}_2(B)$. These do not hold in general (e.g. 9.1.6 or 9.3.5), but they do hold whenever the joint prevision is the independent natural extension (or type-1 product) of its marginals; see 9.3.5.

3. The factorization property holds whenever the first experiment is irrelevant to the second, so that $\underline{P}(Z|\mathcal{X}_1) = \underline{P}_2(Z)$. This is not sufficient for independence in general, because the second experiment may be relevant to the first. See Example 9.2.4. It is sufficient when all outcomes have positive probability, because P determines the conditional probabilities through Bayes' rule.

4. This is Example 4.6 of Billingsley (1979), who defends the standard definition of independence.

5. This is discussed by de Finetti (1972, p. 98), who suggests that we can choose two integers 'at random... and independently', and moreover that this defines a unique joint prevision. It is not clear what model he intends, e.g. what probability should be assigned to the event that $\theta \leq x$?

6. Assuming that the joint prevision P is linear and the marginal previsions have densities f_1 and f_2 with respect to some σ -finite measures, necessary and sufficient conditions for the independence condition 9.2.1 are that:

- (a) P has a joint density f that factorizes, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$; and
- (b) for all $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, $P(\cdot|x_1)$ and $P(\cdot|x_2)$ have densities f_2 and f_1 respectively.

7. De Finetti (1974, section 6.9.4) gives a further example in which there is logical dependence, but independence under the standard definition. Fine (1973, pp. 82, 141–6) has suggested that the factorization property is necessary but not sufficient to characterize physical independence of experiments.

8. On the theory of conditional independence for precise probabilities, see Dawid (1979).

Section 9.3

1. Define $\underline{P}(X|x_1) = \underline{P}_2(X(x_1, \cdot))$ and $\underline{P}(X) = \underline{P}_1(\underline{P}(X|\mathcal{X}_1))$ for $X \in \mathcal{F}$. Then \underline{P} has marginals \underline{P}_1 and \underline{P}_2 . Under this model x_1 is irrelevant to x_2 and the factorization property holds, but the marginals need not be independent (see 9.2.4). For the comparative probability models studied by Kaplan and Fine (1977), joint models cannot always be constructed from given marginals, even if we do not require independence.
2. This is the lower envelope of all joint previsions that are coherent with $\underline{P}_2(\cdot|\mathcal{X}_1)$ and $\underline{P}_1(\cdot|\mathcal{X}_2)$.
3. The condition for compatibility fails for $Y = A - 0.5$, where A is the event that $\theta \leq x$.
4. This is essentially Fubini's formula, permitting a change in the order of integration. For the theory of product measures, see Billingsley (1979), Halmos (1950), Royden (1963) or Taylor (1973).
5. Provided one of the marginal spaces $\mathcal{X}_1, \mathcal{X}_2$ is finite, any two marginals \underline{P}_1 and \underline{P}_2 are compatible. (The compatibility condition can be verified for linear marginals, as we can change the order of integration and summation over a finite set.)
6. When one marginal (\underline{P}_2) is linear, the unique product prevision is the independent lower envelope of products $P_1 \times P_2$, over P_1 in $\mathcal{M}(\underline{P}_1)$. When \underline{P} is an independent lower envelope with non-linear marginals \underline{P}_1 and \underline{P}_2 , it is not necessary for $\mathcal{M}(\underline{P})$ to contain all products $P_1 \times P_2$ for which P_j is in $\mathcal{M}(\underline{P}_j)$.
7. This kind of model is reasonable also when the chances on each trial are precise but known only approximately. Suppose You believe that a coin has a small bias in favour of 'tails', but the exact bias is unknown. Under a maximally precise model, two tosses of the coin would not be epistemically independent for You, because the outcome of one toss would provide information about the bias and have a slight effect on Your beliefs about the other toss. But You could adopt a less precise model, in which the lower and upper probabilities of 'heads' on each toss are 0.4 and 0.5, and the tosses are epistemically independent. (This would be reasonable if You judge that posterior probabilities would lie in this interval, whatever the outcome of the first toss.) More generally, it may often be convenient to replace a model in which there is a small degree of dependence between trials by a less precise (but simpler) model in which there is exact independence. See also note 12.
8. The independent natural extension can always be constructed in this way, by solving finite sets of linear inequalities, provided the marginals \underline{P}_1 and \underline{P}_2 are finitely generated (4.2.1).
9. Clearly $\underline{Q}(A \times B) = \underline{P}_1(A)\underline{P}_2(B)$. Any product prevision satisfies

- $\underline{P}(A \times B) \geq \underline{P}_1(A)\underline{P}_2(B)$, by note 2 of section 9.2. Since the independent natural extension is the minimal product prevision with its marginals, it satisfies $\underline{E}(A \times B) = \underline{P}_1(A)\underline{P}_2(B)$.
10. A different model is obtained if P is known to be an extreme point of $\mathcal{M}(\underline{P})$. In Example 9.3.4, this gives the independent lower envelope of $\{\underline{P}_1, \underline{P}_2\}$ as the product prevision.
 11. The product prevision is unique when one marginal is linear. Compare with earlier results of this kind, that certain extensions are uniquely determined when previsions are linear, but not when they are non-linear. In particular, a joint prevision is not uniquely determined by marginal and conditional previsions (6.7), posterior previsions are not uniquely determined by prior previsions and sampling model (8.5), and previsions are not uniquely determined by probabilities (2.7).
 12. Although these models are based on an assumption of exact independence of precise probabilities, they are consistent with a small degree of dependence. That is because independence is not preserved under convex combinations of linear previsions, so $\mathcal{M}(\underline{P})$ must contain linear previsions that are not product previsions, even when \underline{P} is an independent lower envelope. We can measure the degree of dependence between two binary experiments under P by $\beta(P) = P(B_1 \cap B_2)/P(B_1)P(B_2) - 1$, where B_j denotes success on trial j . Then $\beta = 0$ represents exact independence. The type-1 and type-2 product previsions in Example 9.3.4 are dominated by $P = (\underline{P}_1 + \underline{P}_2)/2$, which has a small degree of dependence $\beta(P) = 0.012$. The independent natural extension \underline{E} is consistent with a much larger degree of dependence, as $\mathcal{M}(\underline{E})$ has extreme points P_5 and P_6 with $\beta(P_5) = 0.125$, $\beta(P_6) = -0.12$. A judgement that imprecise marginals are independent is therefore *robust*, in the sense that it is consistent with some dependence between underlying precise marginals. For that reason, judgements of exact independence seem more reasonable than precise judgements of probabilities.
 13. See Walley and Fine (1982) for further properties of the type-1 and type-2 products. It is not clear what type of product prevision is most suitable for modelling physically independent repetitions of an experiment that is physically indeterminate (7.2.9).

Section 9.4

1. Johnson (1924, p. 183) and (1932), and Good (1965), used the term 'permutability' for the Bayesian concept of exchangeability. Since our definitions of the two terms agree in the case of additive probabilities, our terminology is consistent with Johnson's.
2. Compare with de Finetti (1937, p. 120).

3. See 9.5.2 for further discussion of these examples in connection with exchangeability. Mallows and Walley (1980, section 2.3.3) suggest that unformalized judgements of permutability play an important role in statistical data analysis. What is often called a 'population' in data analysis can be regarded as a set of permutable individuals or units. Permutability judgements justify 'pooling' the units, so that conclusions drawn from the data are invariant under permutations of them. Of course the data themselves may suggest differences between 'subpopulations'. That is consistent with an initial judgement of permutability, but not with a stronger judgement of exchangeability.
4. One difference is that permutability is defined solely in terms of an unconditional prevision. It may be necessary to strengthen the definition by requiring also that conditional previsions be permutation-invariant. That is automatic if all outcomes have positive lower probability.
5. A method for constructing the more precise model is given in 9.7.3.
6. An example is the lower envelope of $\{P_1, P_2, P_3\}$ in 9.3.4. This has independent identical marginals, but it is not permutable because $P(1, 0) = 0.2$ and $P(0, 1) = 0.24$.
7. If \underline{P} is permutable, there are permutable linear previsions in $\mathcal{M}(\underline{P})$. They can all be constructed as follows. For each P in $\mathcal{M}(\underline{P})$, define P^* to be the mean of πP over all permutations π . Then $P^* \in \mathcal{M}(\underline{P})$ by convexity, and P^* is permutable.
8. Because \underline{P} is permutable if and only if $\pi \underline{P} = \underline{P}$, and $P \in \mathcal{M}(\underline{P})$ if and only if $\pi P \in \mathcal{M}(\pi \underline{P})$.

Section 9.5

1. See especially de Finetti (1937, 1975, Ch. 11), (1972, pp. 209–25). Other discussions include Good (1965), Kyburg (1974a), Lindley and Novick (1981), Dawid (1982a, 1982b, 1982c). On the mathematical aspects of exchangeability, see also Hewitt and Savage (1955), Heath and Sudderth (1976), Kingman (1978) and Jaynes (1986).
2. In terms of the almost-preference ordering induced by \underline{P} (see 3.8.1), X and πX are equivalent. That is, each one is almost-preferred to the other.
3. An example is the Bernoulli model 9.3.4. Let A be the event that there is 1 on the first trial and 0 on the second, and let B be the permuted event. The independent natural extension \underline{E} is permutable but not exchangeable, as $\underline{E}(B - A) = -0.1$, $\bar{E}(B - A) = 0.1$. The same is true of the type-1 product prevision.
4. For example, type-2 product previsions (9.3.5) are exchangeable, because they are lower envelopes of linear product previsions with identical marginals.

5. According to de Finetti (1937, p. 131), 'it suffices that the conditions... do not present any apparent asymmetry'. See also Smith (1984, p. 251).
6. The difference between exchangeability and permutability is analogous to the difference between a uniform distribution and the vacuous model, discussed in 5.5.1.
7. It would also be justified if You knew only that the group is a random sample of individuals chosen from the same population, because then the models in 9.5.3 and 9.5.4 are justified. Permutability (but not exchangeability) would be appropriate if You had no information about how the sample was selected. Knowledge of the sampling mechanism therefore supports a more precise probability model, exchangeable rather than permutable. That is one way in which random sampling can be useful.
8. For details, see Heath and Sudderth (1976), de Finetti (1972) and Hewitt and Savage (1955). Heath and Sudderth give a simple proof.
9. Modifications of the standard urn process lead to models that are permutable but not exchangeable. For instance: (a) the balls in the urn are of different sizes, with different chances of being drawn that are not precisely known; (b) the marks change in time, or according to the results of earlier drawings, as in de Finetti (1975, section 11.3.4); or (c) there are several urns of different composition and You have no information about which will be used on any trial.
10. Heath and Sudderth (1976) give a simple and insightful proof. For other proofs, see de Finetti (1937, p. 128), Hewitt and Savage (1955), Good (1965) and Feller (1966, section 7.4). Some generalizations and extensions of the theorem are in Hewitt and Savage (1955), Dawid (1982a, 1982b, 1982c) and Jaynes (1986).
11. Here P_0 can be taken to be countably additive and is then uniquely determined by P .
12. These are respectively the limiting frequency and propensity interpretations of θ , discussed in 7.2. De Finetti (1975, p. 227) strongly rejects both interpretations. For him, θ is merely a parameter that appears in a mathematical representation of exchangeable probabilities.
13. By note 11, P_0 can be taken to be a lower envelope of countably additive, linear previsions.
14. The chances must be precise. Models for independent repetitions of trials with imprecise chances (7.2.9) would be permutable but not exchangeable. The robust Bernoulli model (9.6) can be interpreted in this way.
15. This is discussed by Fine (1970, 1973) and Walley and Fine (1982).
16. Here δ could be taken to be a slowly decreasing function of k .
17. See de Finetti (1937, p. 124), Kyburg (1974a, pp. 122–6) or Dawid (1982c). In effect, You can estimate θ with arbitrary precision as $k \rightarrow \infty$. Cox

and Hinkley (1974, 1978, Problem 10.6) point out that the assumption that a large number of unrelated events are epistemically independent and have approximately the same precise probability p leads to a similar conclusion: You should be practically certain that the relative frequency of occurrence of the events will be close to p .

18. See Walley and Fine (1982, Thm. 4.1). When δ is larger than the marginal imprecision of the model, the lower probability of $C(k)$ under the type-1 product tends to 1 as $k \rightarrow \infty$. Thus the model is consistent with long-run fluctuations in relative frequencies, but these can be no greater than the imprecision of the marginals. These results apply to the robust Bernoulli models (9.6.3).

19. When \underline{P} is permutable, $\underline{P}(C(k))$ can tend to any limit between 0 and 1. (Consider linear-vacuous mixtures where the linear prevision is exchangeable.) The limit can be 1 even if \underline{P} is not exchangeable. Compare with the stationary lower probability models developed by Kumar and Fine (1985), Grize and Fine (1987), Papamarcou (1987) and Fine (1988), in which convergence of relative frequencies has upper probability zero. The upper and lower probabilities in these models are 3-coherent (Appendix B), but they incur sure loss.

20. De Finetti (1937, 1975), Dawid (1982a, 1982b, 1986), Goldstein (1981, 1986). For discussion, see Hacking (1965), Good (1965), Levi (1967) and Kyburg (1974a).

21. De Finetti (1937, pp. 141–2) makes a distinction between drawings from an urn of unknown composition and tosses of a coin. He believes that ‘unknown chances’ are meaningful in the first case but not in the second. The chances for the urn can (in principle) be measured by examining the composition of the urn, whereas the chances for the coin cannot be deduced from its physical characteristics. (What if the coin is physically symmetric?)

22. De Finetti (1937, p. 141) and (1974, pp. 158, 221).

23. This issue is discussed in 2.11.9.

24. The need for aleatory models is discussed in 2.11.2 and 7.2.

25. Here ‘prediction’ and ‘explanation’ are probabilistic, not deterministic. Inference about statistical parameters has perhaps been over-emphasized in the theory of statistics, and too little attention has been given to predictive previsions concerning observables. Nevertheless, inferences about parameters are often useful.

26. For example, two events A_1 and A_2 are exchangeable under the linear prevision determined by $P(A_1 \cap A_2^c) = P(A_1^c \cap A_2) = \frac{1}{2}$. This is not a mixture of Bernoulli models as these satisfy $P(A_1 \cap A_2^c | \theta) = \theta(1 - \theta) \leq \frac{1}{4}$. Here A_1 and A_2 are negatively dependent. It follows from de Finetti’s theorem that when A_1 and A_2 are embedded in an infinite exchangeable sequence, they must be non-negatively dependent (9.7.2). For other results concerning finite exchangeability, see Jaynes (1986).

27. Especially as different types of observables may be related through the same parameter θ . Let θ denote the half-life of a radioactive element. There are many different observables, e.g. the numbers of disintegrations of various masses in various time periods, whose probability distribution depends on θ . It would be foolish to assess predictive probabilities for the observables directly, without referring to θ . Parametric sampling models that are much more complex than the Bernoulli model are often needed in statistical analysis. To replace these by models which refer only to observables can become hopelessly complicated.

Section 9.6

1. It is necessary to define ‘regular trials’, in order that θ is well defined. It is not possible to define θ as the index of the ‘true model’ \mathcal{M}_θ , where \mathcal{M}_θ is a neighbourhood of the standard Bernoulli (θ) model, because the sets \mathcal{M}_θ are not disjoint and the ‘true’ \mathcal{M}_θ is not well determined. (See 7.2.7 for discussion.) There are similar problems with the second interpretation. (It is unrealistic to assume that the upper and lower bounds are sharp.) In any case, θ is not identifiable from repeated observations (9.6.5).

2. The marginal considered in 9.3.4 is an example of this, defined by the values $\theta = \frac{4}{9}$ and $\delta = 0.1$.

3. More complicated, but perhaps more realistic, models can be developed by assessing non-vacuous beliefs about ψ , and non-degenerate beliefs about δ . But if there is little prior information about how the physical conditions vary with time, it is unrealistic to assess a precise prior distribution for ψ . (Compare with the ‘robust’ Bayesian models of Box (1979, 1980), which require precise prior distributions for nuisance parameters such as ψ .) It is possible to learn about θ from the observations, despite the vacuous prior for ψ . That is essentially because the effect of ψ on probabilities is bounded by $\tilde{\delta}$, so that $\underline{P}(1|\theta)$ and $\bar{P}(1|\theta)$ are non-vacuous. It would not be possible to allow vacuous beliefs about δ .

4. This is defined by $\underline{P}(Y) = \inf \{ \underline{P}_0(Y(\cdot, \psi)) : 0 \leq \psi_j \leq 1 \text{ for } 1 \leq j \leq n \}$. It agrees with the type-1 product of \underline{P}_0 and the vacuous prior (9.3.5).

5. This differs from the independent natural extension of these marginals (as in 9.3.4), although that yields the same values of $\underline{P}(x|\theta)$ and $\bar{P}(x|\theta)$. The assumption that underlying chances are precise is a substantive one, even if nothing is known about the values ψ_j . It is this assumption that justifies the type-1 product. Some other product might be more appropriate if the chances were intrinsically imprecise, due to physical indeterminacy in the experiment.

6. Let B_j denote 1 on trial j . Then $\underline{P}(B_1 - B_2) = -\delta$ and $\bar{P}(B_1 - B_2) = \delta$, contrary to the definition of exchangeability.

7. Inferences depend on the experimental evidence only through the precise

likelihood function $P(\underline{x}|\theta, \psi)$. So any two experiments which yield the same outcomes \underline{x} , such as binomial and negative binomial experiments, will lead to the same inferences about θ . This is an application of the discrete likelihood principle (8.6.1).

8. This is an imprecise beta prior of the type defined in 5.3.1, with learning parameter $s_0 = 4$.

9. We assume that $\delta \leq r \leq 1 - \delta$. Otherwise, approximate the upper or lower prevision by one or zero.

Section 9.7

1. These constraints are sufficient for independence provided B has positive lower probability, but not in general, because \mathcal{E} may not determine conditional revisions (3.8.6, 9.2). The constraints would be sufficient in general if we took \mathcal{E} to be a class of really desirable (rather than almost-desirable) gambles, as in Appendix F. This comment also applies to judgements of types (c)–(f).

2. See Appendix G and note 6 of section 4.1. Similarly, a judgement that $\underline{V}(X) \leq 3$ is not a structural judgement. But $\underline{V}(X) \geq 3$ is a direct judgement, and $\bar{V}(X) \geq 3$ is a structural judgement. (The latter is equivalent to the judgements that $\lambda - (X - \mu)^2$ is nondesirable, for all real μ and all $\lambda < 3$.)

3. The key property of structural constraints is that they are preserved under intersections of classes \mathcal{E} . This ensures the existence of a minimal coherent \mathcal{E} that satisfies the structural constraints. More generally, we could allow \mathcal{J} to include any constraints on \mathcal{E} that are preserved under intersections. Alternatively, specify constraints on unconditional and conditional lower revisions that are preserved under forming lower envelopes, and apply 7.1.6 to establish coherence.

4. So \underline{P} is the lower envelope of all coherent lower revisions \underline{Q} that satisfy $\underline{Q}(X) \geq 0$ whenever $X \in \mathcal{D}$, and $\underline{Q}(Y) \geq 0$ whenever $(X, Y) \in \mathcal{J}$ and $\underline{Q}(X) \geq 0$.

5. This does not always work, because $\mathcal{E}(\mathcal{D}^*)$ does not always satisfy the structural constraints.

6. It cannot be determined from \underline{P} alone which models \underline{Q} are consistent with the structural judgements. This can cause difficulties in the type of sequential assessment described in 4.3. The new model \underline{P} constructed at each stage will depend not only on the previous model, but also on what structural judgements were made at earlier stages.

7. For further discussion, see 2.10.3.

8. The example of indeterminacy is simplest. If the gambles $A - 0.45$ and $0.55 - A$ are indeterminate for You, so You have no disposition to accept either one, then Your dispositions cannot be modelled by any linear prevision. For independence and permutability, see 9.1.7 and 9.4.5.

9. See 9.1.5 and 9.3.5. Sensitivity analysis has the advantage that models such as type-1 products are easier to construct than independent natural extensions (e.g. 9.3.4). For both products, the number of extreme points increases very quickly with the number of marginals, but the extreme points of the type-1 product are more tractable because they are all products of linear previsions.

APPENDIX A

Verifying coherence

The general conditions for avoiding sure loss (2.4.1, 2.4.4) and coherence (2.5.1, 2.5.4) can be difficult to verify in practice, especially when either the possibility space Ω or the domain \mathcal{K} is large. The following results help to reduce the computational effort needed to check whether a numerically specified lower prevision is coherent or avoids sure loss.

A1 Lemma

In verifying the avoiding sure loss condition 2.4.4(a) or the coherence condition 2.5.4(a), it suffices to consider positive values of $\lambda_1, \dots, \lambda_n$ and gambles $G(X_1), \dots, G(X_n)$ that are linearly independent in the linear space $\mathcal{L}(\Omega)$. In verifying coherence, when $\lambda_0 > 0$ we can also assume that $G(X_0), \dots, G(X_n)$ are linearly independent.

Proof. For 2.4.4(a), suppose that \underline{P} incurs sure loss, so there are $X_j \in \mathcal{K}$ and $\lambda_j \geq 0$ such that $\sup \sum_{j=1}^n \lambda_j G(X_j) < 0$. Let n be minimal with this property, so clearly $\lambda_1, \dots, \lambda_n$ are all positive. Suppose $G(X_1), \dots, G(X_n)$ are linearly dependent, i.e. there are μ_1, \dots, μ_n (not all zero) such that $\sum_{j=1}^n \mu_j G(X_j) = 0$. Consider

$$\sup \sum_{j=1}^n (\lambda_j + c\mu_j) G(X_j) = \sup \sum_{j=1}^n \lambda_j G(X_j) < 0, \text{ for any real } c.$$

We can choose c so that $\lambda_j + c\mu_j \geq 0$ for $1 \leq j \leq n$, with $\lambda_j + c\mu_j = 0$ for some j . (If some $\mu_i > 0$, choose $c = -\min\{\lambda_j/\mu_j; \mu_j > 0\}$. If all $\mu_i \leq 0$ then choose $c = \min\{-\lambda_j/\mu_j; \mu_j < 0\}$.) Thus some $G(X_j)$ can be removed to reduce n , contrary to the assumption of minimality.

For 2.5.4(a), suppose \underline{P} is not coherent, so $\sup[\sum_{j=1}^n \lambda_j G(X_j) - \lambda_0 G(X_0)] < 0$. Let n be minimal with this property. If $\lambda_0 = 0$, \underline{P} incurs sure loss and the first result applies. Otherwise, we must have $\lambda_0, \lambda_1, \dots, \lambda_n$ all positive. If $G(X_0), \dots, G(X_n)$ are linearly dependent, again we can remove one of them. If X_0 is removed, the remaining gambles incur sure loss, and otherwise n is reduced, contrary to minimality. ♦

A2 Theorem

Suppose Ω is finite, with n elements, and the lower prevision \underline{P} on \mathcal{K} satisfies the **non-negativity** condition: $\{\omega\} \in \mathcal{K}$ and $\underline{P}(\{\omega\}) \geq 0$ for all $\omega \in \Omega$.

(a) \underline{P} avoids sure loss if and only if there is no linearly independent set $G(X_1), \dots, G(X_m)$, where $1 \leq m \leq n$ and all $X_j \in \mathcal{K}$, such that $1, G(X_1), \dots, G(X_m)$ is a linearly dependent set, and the unique values $\lambda_1, \dots, \lambda_m$ such that $\sum_{j=1}^m \lambda_j G(X_j) = -1$ are all positive.

(b) \underline{P} is coherent if and only if it avoids sure loss and there is no linearly independent set $G(X_0), G(X_1), \dots, G(X_m)$, where $0 \leq m \leq n-1$ and all $X_j \in \mathcal{K}$, such that $1, G(X_0), \dots, G(X_m)$ is a linearly dependent set, and the unique values $\lambda_0, \dots, \lambda_m$ such that $\sum_{j=1}^m \lambda_j G(X_j) - \lambda_0 G(X_0) = -1$ are all positive.

Proof. (a) If such positive λ_j exist, \underline{P} incurs sure loss. Conversely, suppose \underline{P} incurs sure loss, so $\sup \sum_{j=1}^k \lambda_j G(X_j) < 0$ for some $X_j \in \mathcal{K}$, $\lambda_j > 0$. By the non-negativity condition, we can add new gambles $X_j = \{\omega_j\}$ to give $\sum_{j=1}^m \lambda_j G(X_j) = -c < 0$. By renormalizing λ_j we can assume $c = 1$. By the argument of Lemma A1 we can further assume that $G(X_1), \dots, G(X_m)$ are linearly independent, so that $\lambda_1, \dots, \lambda_m$ are uniquely determined.

(b) If such positive λ_j exist, \underline{P} is not coherent. Conversely, suppose \underline{P} avoids sure loss but is not coherent. Using non-negativity and renormalizing the λ_j we can find $X_0, \dots, X_m \in \mathcal{K}$ and $\lambda_j > 0$ such that $\sum_{j=1}^m \lambda_j G(X_j) - \lambda_0 G(X_0) = -1$. Again the gambles $G(X_0), \dots, G(X_m)$ can be reduced to a linearly independent set, so that the values λ_j are uniquely determined. ♦

The theorem suggests a general algorithm for checking whether a given lower prevision is coherent. To check the conditions of the theorem it is necessary, for each linearly independent set of gambles, to solve n linear equations for the (at most n) unknowns λ_j . Essentially, this requires inverting matrices of size up to $n \times n$. Whether the conditions can be checked in practice will depend on n (the size of Ω) and on the size of \mathcal{K} , which affects the number of linearly independent sets of gambles. The dual problem, which involves finding the linear previsions that dominate \underline{P} , is important for theoretical reasons and may often be computationally simpler. It is described in section 4.2.

The assumption of non-negativity in (a) of the theorem is not substantive, because if $\underline{P}(\{\omega\})$ is not defined or is negative for some ω we can modify \underline{P} by setting $\underline{P}(\{\omega\}) = 0$ without affecting whether \underline{P} avoids sure loss. In the case of coherence, the non-negativity assumption is a substantive restriction on the domain \mathcal{K} . But if \mathcal{K} already contains all gambles $\{\omega\}$ then non-negativity is necessary for coherence.

To check whether a given lower probability is coherent, we can use the following modification of Theorem A2. (This can be proved using the ideas in the two previous proofs.)

A3 Theorem

Suppose Ω is finite, with n elements, and \underline{P} is a lower probability defined on \mathcal{A} which satisfies $\underline{P}(\emptyset) = 0$, $\underline{P}(\Omega) = 1$, $\{\omega\} \in \mathcal{A}$ and $\underline{P}(\{\omega\}) \geq 0$ for all $\omega \in \Omega$.

- (a) \underline{P} avoids sure loss if and only if there is no linearly independent set of events A_1, \dots, A_m in \mathcal{A} , where $2 \leq m \leq n$, such that the unique values $\lambda_1, \dots, \lambda_m$ satisfying $\sum_{j=1}^m \lambda_j A_j = 1$ are all positive and $\sum_{j=1}^m \lambda_j \underline{P}(A_j) > 1$.
- (b) \underline{P} is coherent if and only if

- (i) it avoids sure loss;
- (ii) there is no non-trivial event A_0 and linearly independent set of events A_1, \dots, A_m in \mathcal{A} , where $2 \leq m \leq n-1$, such that the unique values $\lambda_1, \dots, \lambda_m$ satisfying $\sum_{j=1}^m \lambda_j A_j = A_0$ are all positive and $\sum_{j=1}^m \lambda_j \underline{P}(A_j) > \underline{P}(A_0)$;
- (iii) there is no linearly independent set of events A_0, A_1, \dots, A_m in \mathcal{A} , where $2 \leq m \leq n-1$, such that the unique values $\lambda_0, \dots, \lambda_m$ satisfying $\sum_{j=1}^m \lambda_j A_j - \lambda_0 A_0 = 1$ are all positive and $\sum_{j=1}^m \lambda_j \underline{P}(A_j) - \lambda_0 \underline{P}(A_0) > 1$.

APPENDIX B

n-coherence

The general definitions of avoiding sure loss (2.4.1) and coherence (2.5.1) involve n gambles, where n is an arbitrary positive integer. Violations of these conditions for small values of n seem more serious than violations for large n . We can formalize this idea by restricting the avoiding sure loss and coherence conditions to hold only when they involve no more than a fixed number of gambles, as in the following definitions.

B1 Definition

Let \underline{P} be a lower prevision on domain \mathcal{K} .

- (a) Say that \underline{P} is *n-asl* if $\sup \sum_{j=1}^n \lambda_j G(X_j) \geq 0$ whenever $X_1, \dots, X_n \in \mathcal{K}$ and each $\lambda_j \geq 0$. (Here *asl* is an abbreviation for ‘avoiding sure loss’.)
- (b) Say that \underline{P} is *n-coherent* if it is *n-asl* and also $\sup [\sum_{j=1}^{n-1} \lambda_j G(X_j) - \lambda_0 G(X_0)] \geq 0$ whenever $X_0, X_1, \dots, X_{n-1} \in \mathcal{K}$, each $\lambda_j \geq 0$ and $\lambda_0 > 0$.

By Lemmas 2.4.4 and 2.5.4, \underline{P} avoids sure loss if and only if it is *n-asl* for all $n \geq 1$, and \underline{P} is coherent if and only if it is *n-coherent* for all $n \geq 1$. The property of *n*-coherence clearly increases in strength as n increases, but it is never stronger than coherence. So *n*-coherence defines a sort of *bounded rationality*, in that it requires Your previsions to be consistent only when restricted to small domains, containing no more than n gambles.

If either of the sets Ω or \mathcal{K} is finite, let n be the cardinality of the smaller set. Then *n*-coherence is equivalent to coherence and *n-asl* is equivalent to avoiding sure loss. (That is clear from the definitions when $|\mathcal{K}| = n$, and follows from Theorem A2 when $|\Omega| = n$ since then the linear space $\mathcal{L}(\Omega)$ has dimension n .)

The basic consequences of avoiding sure loss, (a)–(g) of 2.4.7, are consequences of 3-asl. Similarly the basic properties 2.6.1 (a)–(l) of coherent lower previsions are consequences of 4-coherence. Because the coherence axioms P1–P3 are consequences of 3-coherence, we can characterize *n*-coherence on linear spaces in terms of the following simple axioms.

B2 Theorem

Suppose \underline{P} is defined on a linear space \mathcal{K} .

- (a) \underline{P} is 1-coherent if and only if $\inf X \leq \underline{P}(X) \leq \sup X$ whenever $X \in \mathcal{K}$.
- (b) \underline{P} is 2-coherent if and only if it satisfies axioms P2, P4 and the condition: $\underline{P}(X) \leq -\underline{P}(-X)$ whenever $X \in \mathcal{K}$.
- (c) For $n \geq 3$, \underline{P} is n -coherent if and only if it satisfies axioms P1, P2, P3. (Then n -coherence is equivalent to coherence.)

(These results hold under the weaker assumption that \mathcal{K} is a convex cone, provided P1 is replaced by P4 in (c), and the last condition in (b) is replaced by the stronger condition: $\underline{P}(X) + \underline{P}(Y) \leq \mu$ whenever $X + Y \leq \mu$.)

The next theorem characterizes n -coherence of lower probabilities defined on a field, when $n \leq 3$. It can be proved by using Theorem A3.

B3 Theorem

Suppose \underline{P} is defined on a field of events \mathcal{A} . Let \bar{P} be the conjugate upper probability, $\bar{P}(A) = 1 - \underline{P}(A^c)$.

- (a) \underline{P} is 1-coherent if and only if $\underline{P}(\Omega) = 1$, $\underline{P}(\emptyset) = 0$ and $0 \leq \underline{P}(A) \leq 1$ for all $A \in \mathcal{A}$.
- (b) \underline{P} is 2-coherent if and only if it is 1-coherent and also satisfies $\underline{P}(A) \leq \bar{P}(A)$ and $\underline{P}(A) \leq \underline{P}(B)$ whenever $B \supseteq A$.
- (c) \underline{P} is 3-coherent if and only if it is 1-coherent and satisfies $\bar{P}(A \cup B) \leq \bar{P}(A) + \bar{P}(B)$ and $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B)$ whenever $A \cap B = \emptyset$. (An equivalent condition is that $\underline{P}(A) + \underline{P}(B) + \bar{P}(C) \leq 1 \leq \underline{P}(A) + \bar{P}(B) + \bar{P}(C)$ whenever $\{A, B, C\}$ is a partition of Ω .)

Some authors call a set function \underline{P} a lower probability only when it satisfies the 3-coherence conditions in (c); see Suppes (1974), Wolfson and Fine (1982) and Fine (1983). The 3-coherence conditions are sufficient for coherence when Ω has only 3 elements (as in many examples in Chapter 4), but not when Ω has 4 or more elements. Example 2.7.6, with 4 elements, satisfies 3-coherence but not 4-coherence. Indeed, Example 2.7.5 shows that 3-coherence is not sufficient even for avoiding sure loss.

B4 Comparative probability

The definition of n -coherence can be applied to the comparative probability orderings discussed in section 4.5. A partial ordering \geqslant is n -asl if, whenever $A_j \geqslant B_j$ and $\lambda_j \geq 0$ for $1 \leq j \leq n$, there is some $\omega \in \Omega$ such that $\sum_{j=1}^n \lambda_j (A_j - B_j)(\omega) \geq 0$. The ordering is n -coherent if it is n -asl and also satisfies $A_0 \geqslant B_0$ whenever $A_0 - B_0 \geq \sum_{j=1}^{n-1} \lambda_j (A_j - B_j)$ with $A_j \geqslant B_j$ and

$\lambda_j \geq 0$ for $1 \leq j \leq n-1$. Thus n -coherence requires consistency amongst sets of no more than n comparisons $A_j \geqslant B_j$.

For small values of n , n -coherence of \geqslant can be characterized in terms of the simple properties listed in section 4.5.3:

1. \geqslant is 1-coherent if and only if it satisfies properties (i) and (vi) of 4.5.3;
2. \geqslant is 2-coherent if and only if it satisfies (ii), (iii), (iv), (v), (viii) and (ix);
3. \geqslant is 3-coherent if and only if it satisfies (i), (ii), (iii), (viii), (x) and (xiii).

B5 De Finetti's axioms

Most work on comparative probability has been based on five axioms proposed by de Finetti (1931). These axioms are properties (i), (ii), (viii), (x) of 4.5.3, together with the completeness axiom 4.5.6 (for any events A and B , either $A \geqslant B$ or $B \geqslant A$). All complete coherent orderings satisfy de Finetti's axioms. In fact, these axioms are equivalent to completeness plus 3-coherence. (Conditions (iii) and (xiii) for 3-coherence can be derived from the other axioms.)

From our point of view, de Finetti's completeness axiom is unduly restrictive. (It might be replaced, in view of the above characterization of 3-coherence, by the weaker conditions (iii) and (xiii) of 4.5.3.) When completeness holds, de Finetti's axioms are too lenient since they allow incoherence amongst sets of four or more comparisons, as in the following example.

B6 Example (Kraft, Pratt and Seidenberg, 1959)

Let $\Omega = \{a, b, c, d, e\}$, $A_0 = \{b, e\}$, $B_0 = \{a, c, d\}$, $A_1 = \{d\}$, $B_1 = \{a, c\}$, $A_2 = \{b, c\}$, $B_2 = \{a, d\}$, $A_3 = \{a, e\}$ and $B_3 = \{c, d\}$. Kraft, Pratt and Seidenberg constructed a complete ordering of the subsets of Ω which satisfies de Finetti's axioms and is therefore 3-coherent, such that $A_1 > B_1$, $A_2 > B_2$, $A_3 > B_3$ and $B_0 > A_0$. Since $A_0 - B_0 = (A_1 - B_1) + (A_2 - B_2) + (A_3 - B_3)$, this ordering violates 4-coherence. This shows that de Finetti's axioms are strictly weaker than coherence plus completeness.

This ordering does avoid sure loss and it therefore has a coherent natural extension \geqslant^* , in which $A_j \approx^* B_j$ (meaning that both $A_j \geqslant^* B_j$ and $B_j \geqslant^* A_j$) for $j = 0, 1, 2, 3$. Another ordering constructed by Kraft, Pratt and Seidenberg, on a space with six elements, satisfies de Finetti's axioms but incurs sure loss.

APPENDIX C

Win-and-place betting on horses

The three fundamental concepts of avoiding sure loss, coherence and natural extension can be illustrated by the example of win-and-place betting on horses. Consider a horse race involving n horses numbered 1, 2, ..., n . It is usually possible to bet on any horse to **win** or to **place** (i.e., to finish in the first three, called ‘show’ in North America), either with a bookmaker or through a pari-mutuel or totalizator system. Our point of view here is that of a punter trying to exploit any incoherence in the quoted selling prices \bar{P} of a bookmaker or pari-mutuel system. We suggested a model in section 2.9.3 for the case in which only ‘win’ bets are allowed, but win-and-place betting introduces extra complications and raises the possibility (theoretical, at least) of arranging a ‘sure gain’ for the bettor. That is, it may be possible for the punter to place a system of bets that guarantees him a profit, irrespective of the outcome of the race. There is such a system just when the selling prices \bar{P} incur sure loss. Incoherence of \bar{P} also has an interesting interpretation. It occurs just when the quoted betting rates \bar{P} can, in effect, be improved by a combination of bets on other events. In this appendix we give necessary and sufficient conditions for avoiding sure loss and coherence. These generalize the results of Heath and Sudderth (1972), who studied the problem of win-and-place betting under the (unrealistic) assumption that the prices \bar{P} were additive probabilities. Finally, we discuss a simple interpretation of natural extension, which summarizes the minimum prices You must pay to bet on several different horses in the same race.

Let A_i denote the event that horse i wins the race, B_i the event that horse i places, and $\mathcal{A} = \{A_i, B_i : 1 \leq i \leq n\}$ the set of $2n$ events on which one can bet. Note that $B_i \supset A_i$, $\sum_{i=1}^n A_i = 1$ and $\sum_{i=1}^n B_i = 3$. The domain \mathcal{A} is not closed under unions or intersections of events. In effect, a bookmaker or pari-mutuel system specifies an upper probability \bar{P} on the domain \mathcal{A} . For each event $C \in \mathcal{A}$, one can effectively buy the gamble C for price $\bar{P}(C)$. (This model does not apply to North America, where the payout when horse i places depends on what other horses place, but does apply to most other countries. See Epstein, 1977.) In practice, betting rates are measured in other

ways, by the payout (or reward) for a bet of 1 unit on C , or by the odds against C . When $\bar{P}(C)$ is the upper probability, the payout is $\bar{P}(C)^{-1}$, and the odds against C are $\bar{P}(C)^{-1} - 1$ to 1. When $\bar{P}(C) = 0.2$, for instance, the payout is 5 units for each unit bet, and the odds are 4 to 1 against. To simplify the notation we write $a_i = \bar{P}(A_i)$ and $b_i = \bar{P}(B_i)$.

The following result characterizes the conditions under which a punter can place a system of bets to guarantee a ‘sure gain’.

C1 *Sure gain theorem*

The upper probability \bar{P} on \mathcal{A} avoids sure loss (so there is no sure gain for a bettor) if and only if it satisfies

- (a) $a_i \geq 0, b_i \geq 0$ for $1 \leq i \leq n$
- (b) $\sum_{i=1}^n \min\{a_i, b_i\} \geq 1$
- (c) $\sum_{i=1}^n \min\{b_i, 1\} \geq 3$.

Proof. Write $\alpha_i = \min\{a_i, b_i\}$ and $\beta_i = \min\{b_i, 1\}$. If any of the conditions fail, the following betting systems guarantee a sure gain for the bettor. If $a_i < 0$ or $b_i < 0$, bet on A_i or B_i respectively. If (b) fails, for each i bet α_i on whichever of A_i and B_i has smaller price, yielding a sure gain of at least $1 - \sum_{i=1}^n \alpha_i$. If (c) fails, bet b_i on each B_i for which $\bar{P}(B_i) < 1$, yielding a sure gain of at least $3 - \sum_{i=1}^n \beta_i$.

Conversely, suppose \bar{P} satisfies the three conditions. Let $\varepsilon_i = \min\{a_i, b_i, 1\}$. Then β_i and ε_i satisfy the constraints $0 \leq \varepsilon_i \leq \beta_i \leq 1$, $\sum_{i=1}^n \beta_i \geq 3$ and $\sum_{i=1}^n \varepsilon_i \geq 1$. Consider $\sum_{i=1}^n (c_i \varepsilon_i + d_i \beta_i)$ as a linear function of the $2n$ variables β_i and ε_i , where c_i and d_i are fixed non-negative integers for $1 \leq i \leq n$. We seek to minimize this function subject to the given constraints on β_i and ε_i . Since $c_i, d_i \geq 0$, the function is minimized when $\sum_{i=1}^n \beta_i = 3$ and $\sum_{i=1}^n \varepsilon_i = 1$. The constraints then define a compact convex set in \mathbb{R}^{2n} , so the minimum is achieved by an extreme point of this set. The extreme points satisfy maximal sets of equality constraints and hence have the form $\varepsilon_i = \beta_i = \beta_j = \beta_k = 1$ for some distinct integers $i, j, k \in \{1, \dots, n\}$ and $\varepsilon_m = \beta_m = 0$ otherwise. The minimum value of the linear function subject to the constraints is therefore $\min\{c_i + d_i + d_j + d_k : i, j, k \text{ distinct}\}$. But if we represent the outcome of the race by $\omega = (i, j, k)$ denoting the first three horses in order, $c_i + d_i + d_j + d_k = \sum_{m=1}^n [c_m A_m(\omega) + d_m B_m(\omega)]$ is the reward for bets of $c_m a_m$ on each A_m and $d_m b_m$ on each B_m . Thus we have shown

$$\begin{aligned} \min_{\omega \in \Omega} \sum_{m=1}^n [c_m A_m(\omega) + d_m B_m(\omega)] &= \min\{c_i + d_i + d_j + d_k : i, j, k \text{ distinct}\} \\ &\leq \sum_{m=1}^n (c_m \varepsilon_m + d_m \beta_m) \leq \sum_{m=1}^n (c_m a_m + d_m b_m). \end{aligned}$$

It follows from section 2.4.6 that \bar{P} avoids sure loss. ◆

C2 Pari-mutuel model

Theorem C1 applies to any system of betting rates. As a special case, consider the pari-mutuel system discussed in section 2.9.3. A fraction τ of the total ‘win’ stakes is deducted, and the remainder is divided amongst those who backed the winner in proportion to their stakes. Similarly, the same fraction τ of the total ‘place’ stakes is deducted, the remainder is divided into three equal parts, and each part is divided amongst those who backed one of the three placed horses in proportion to their stakes. (In 1981, the deduction τ ranged from 0.15 in Australia and Eire, 0.16 in the USA, 0.20 in New Zealand, 0.22 in Great Britain, to 0.40 in Venezuela and 0.46 in Argentina). Ignoring the rounding of payouts, the selling prices for gambles $C \in \mathcal{A}$ under this model are simply $\bar{P}(C) = P_0(C)/(1 - \tau)$, where $P_0(A_i)$ is the fraction of ‘win’ stakes bet on horse i , and $P_0(B_i)$ is the fraction of ‘place’ stakes bet on horse i multiplied by 3. (Of course, any bets made to exploit a ‘sure loss’ will change the effective prices \bar{P} . We ignore this effect here. It will be small provided the stakes of the bets are small, although then the bettor’s ‘sure gain’ will also be small.)

For pari-mutuel prices, condition (a) of Theorem C1 is always satisfied, but there can be a ‘sure gain’ when either (b) or (c) fails. Write $\zeta_i = P_0(A_i) = (1 - \tau)a_i$ and $\eta_i = P_0(B_i) = (1 - \tau)b_i$, so that $\sum_{i=1}^n \zeta_i = 1$ and $\sum_{i=1}^n \eta_i = 3$. Then condition (b) of C1 reduces to

$$(1) \sum_{i: \zeta_i > \eta_i} (\zeta_i - \eta_i) \leq \tau$$

and condition (c) is equivalent to

$$(2) \eta_i \leq 1 + 2\tau \quad \text{for } 1 \leq i \leq n, \text{ and}$$

$$(3) \eta_i + \eta_j \leq 2 + \tau \quad \text{for } 1 \leq i < j \leq n.$$

Thus conditions (1), (2), (3) are necessary and sufficient for the pari-mutuel prices to avoid sure loss.

Note that the prices avoid sure loss unless some $\zeta_i > \eta_i$ (so the fraction of win stakes bet on i is more than 3 times the fraction of place stakes bet on i), or some $\eta_i > 1$ (so more than $\frac{1}{3}$ of total place stakes are bet on i), and even in these cases \bar{P} may avoid sure loss if the deduction τ is sufficiently large. Consider a typical racetrack deduction of $\tau = 0.2$. Condition (1) is almost always satisfied (it is rare even to have $\zeta_i > \eta_i$). Condition (2) is violated when more than $\frac{7}{15}$ of the place stakes are bet on a single horse. Condition (3) is violated when more than $\frac{11}{15}$ of the place stakes are bet on two horses. When conditions (2) or (3) are violated, a system of place bets on each horse for which $b_i < 1$, with stakes proportional to b_i , will guarantee a sure gain for the bettor.

Sure gains from systems of win-and-place bets are rarely possible in practice. When other kinds of bets are available, however, there are many more possibilities for arranging a sure gain. For example, the New Zealand Totalizator Agency Board offers **quinellas** (predicting the first two horses in either order) and **trifectas** (predicting the first three in the correct order), as well as win-and-place betting and **doubles** or **trebles** involving several races. It seems likely that the betting rates \bar{P} on the many different kinds of events sometimes incur sure loss, although that cannot be exploited at present because pre-race information about quinella or trifecta payouts is unreliable or unavailable. Further possibilities for arranging a sure gain exist when one can place bets with several bookmakers, who may each avoid sure loss but fail to do so as a group because they offer different odds. The general point is that the more probability assessments You make concerning related events, the more likely it is that they incur sure loss. Of course, incoherence is even more likely to appear.

C3 Coherence

Next we characterize coherence of the prices \bar{P} for the general case of win-and-place betting. Assuming that \bar{P} avoids sure loss, \bar{P} is incoherent just when there is a system of win-and-place bets from which the bettor can obtain some gamble $C \in \mathcal{A}$ for a price less than $\bar{P}(C)$. Then a punter who wishes to bet on C can do so indirectly, through other bets, for less than the quoted betting rate $\bar{P}(C)$.

It can be shown that an upper probability \bar{P} on \mathcal{A} is coherent if and only if it satisfies the three conditions:

$$(d) 0 \leq a_i \leq b_i \leq 1 \quad \text{for } 1 \leq i \leq n$$

$$(e) \sum_{i=1}^n a_i \geq 1$$

$$(f) \sum_{i=1}^n b_i \geq 3.$$

These conditions are easily seen to be necessary for coherence. They are equivalent to (a), (b), (c) of Theorem C1 and $a_i \leq b_i \leq 1$ for $1 \leq i \leq n$. The last property is necessary since $\Omega \supset B_i \supset A_i$.

Thus \bar{P} avoids sure loss but is incoherent just when $a_i > b_i$ or $b_i > 1$ for some horse i . In the first case we can improve the quoted win odds for i by backing i for a place. In the second case we can improve the quoted place odds for i by not betting at all.

For the pari-mutuel model in section C2, the coherence conditions (e) and (f) hold automatically, and (d) reduces to $\zeta_i \leq \eta_i \leq 1 - \tau$ for $1 \leq i \leq n$.

If $\tau = 0.2$, for example, the pari-mutuel prices are incoherent whenever more than $\frac{4}{15}$ of the place stakes are bet on a single horse.

C4 Natural extension

When \bar{P} is coherent, a single win or place gamble cannot be bought more cheaply through a complicated system of bets. But a punter might want to back some horse for both win and place, or to back several horses in one race. For example, he might want to back each of horses 1, 2, 3 to place, by buying the gamble $B_1 + B_2 + B_3$. The obvious way to do so is through separate place bets on the three horses, at a cost of $b_1 + b_2 + b_3$, but that is not necessarily the cheapest way, even when \bar{P} is coherent. As an artificial example, consider the pari-mutuel model with $\tau = \frac{1}{6}$, $n = 4$, fractions $(\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{3}{4})$ of the win stakes bet on the four horses, and $\frac{1}{4}$ of the place stakes bet on each horse. The pari-mutuel upper probabilities are then $a_j = 0.1$ for $j = 1, 2, 3$, $a_4 = 0.9$, and $b_j = 0.9$ for each j . These prices are coherent, by the results of section C3. Suppose You want to obtain the gamble $X = B_1 + B_2 + B_3$. To do so by betting on each of 1, 2, 3 to place will cost $b_1 + b_2 + b_3 = 2.7$. You can do better by betting 0.1 on each of 1, 2, 3 to win, which yields the gamble $Y = 2 + A_1 + A_2 + A_3$ for price 2.3. Since $Y = 3 - A_4 \geq 3 - B_4 = X$, the gamble Y is better than X . (In fact, 2.3 is the minimum price for which X , or some better gamble, can be bought. To see that, note that \bar{P} dominates an additive probability P which assigns win probabilities $(0.1, 0.1, 0.1, 0.7)$ and assigns probability zero to the event that horse 4 finishes second or third. The natural extension \bar{E} satisfies $\bar{E}(X) \geq P(X) = 3 - P(B_4) = 2.3$.)

Generally, a punter may want to buy an arbitrary gamble X in $\mathcal{L}(\Omega)$, where Ω is the set of possible outcomes of the race (3-tuples listing the first 3 horses in order). The minimum price he needs to pay to obtain X (or a better gamble) is just $\bar{E}(X)$, where \bar{E} is the natural extension of \bar{P} to a coherent upper prevision on $\mathcal{L}(\Omega)$.

Suppose, for example, that X is in the linear space generated by $\{A_1, \dots, A_n\}$, i.e. the reward of X depends only on which horse wins. Write $X = \sum_{j=1}^n x_j A_j$, re-order the horses so that $x_1 \geq x_2 \geq \dots \geq x_n$, and let m be minimal such that $\sum_{j=1}^m a_j \geq 1$. We can construct an additive probability $P \leq \bar{P}$ such that $P(A_j) = a_j$ for $j < m$ and $P(A_j) = 0$ for $j > m$, and clearly this maximizes $P(X)$ over all $P \leq \bar{P}$. It follows from Theorem 3.4.1 that $\bar{E}(X) = P(X) = x_m + \sum_{j=1}^{m-1} (x_j - x_m) a_j$ is the minimum price for which X can be obtained. (In fact, the better gamble $Y = x_m + \sum_{j=1}^{m-1} (x_j - x_m) A_j$ can be obtained for the same price, by betting $(x_j - x_m) a_j$ on each A_j for $j < m$.) Note that $\bar{E}(X)$ depends only on the values a_j , and not on the b_j . In other

words, provided the selling prices are coherent gambles whose reward depends only on which horse wins can be bought for minimum price through win bets alone. The earlier example shows that gambles whose reward depends only on which horses are placed (and not on which of them wins) cannot always be bought optimally through place bets alone.

APPENDIX D

Topological structure of \mathcal{L} and \mathcal{P}

Many of the mathematical results in the book are based on the theory of linear topological spaces. Although most of our results concerning lower previsions used only the elementary theory of linear spaces, topological ideas were needed to relate lower previsions \underline{P} to their classes $\mathcal{M}(\underline{P})$ of dominating linear previsions, through the separating hyperplane and Krein–Milman theorems stated in Appendix E.

In this appendix we present the basic properties of \mathcal{L} and \mathcal{P} as linear topological spaces. The class \mathcal{L} of all gambles on Ω is a linear topological space under the supremum-norm topology, and the class \mathcal{P} of all linear previsions on \mathcal{L} is a subset of the dual space \mathcal{L}^* . The topology on \mathcal{L}^* that is most useful for our purposes is the weak* topology, under which \mathcal{P} is a compact convex subset of \mathcal{L}^* .

Holmes (1975) is an excellent reference on the theory of linear topological spaces and various versions of the separating hyperplane theorem. See also Berberian (1974), Dunford and Schwartz (1958) or Kelley and Namioka (1963) on linear topological spaces, Day (1973) or Jameson (1974) on normed linear spaces, and Stoer and Witzgall (1970) on finite-dimensional linear spaces.

D1 Linear spaces

A set \mathcal{X} is called a **real linear space** (or **real vector space**) when operations of addition and scalar multiplication are defined on it, so that $x + y \in \mathcal{X}$ and $\lambda x \in \mathcal{X}$ whenever $x \in \mathcal{X}$, $y \in \mathcal{X}$ and $\lambda \in \mathbb{R}$, and \mathcal{X} contains a zero vector 0, with the properties (for all $x, y, z \in \mathcal{X}$ and all $\lambda, \mu \in \mathbb{R}$)

- $$\begin{array}{ll} (1) & x + y = y + x, \\ (2) & x + (y + z) = (x + y) + z, \\ (3) & x + 0 = x, \\ (4) & x + (-x) = 0 \quad \text{for some } -x \in \mathcal{X}, \\ (5) & (\lambda + \mu)x = \lambda x + \mu x, \\ (6) & \lambda(x + y) = \lambda x + \lambda y, \\ (7) & (\lambda\mu)x = \lambda(\mu x), \\ (8) & 1x = x. \end{array}$$

A subset \mathcal{T} of \mathcal{X} is called a **linear subspace** when it is closed under linear combinations, i.e. $\lambda x + \mu y \in \mathcal{T}$ whenever $x, y \in \mathcal{T}$ and $\lambda, \mu \in \mathbb{R}$.

For a non-empty set Ω , the class $\mathcal{L} = \mathcal{L}(\Omega)$ of all gambles (bounded mappings from Ω to \mathbb{R}) forms a linear space under the natural pointwise operations $(X + Y)(\omega) = X(\omega) + Y(\omega)$ and $(\lambda X)(\omega) = \lambda X(\omega)$, with zero vector $0(\omega) = 0$.

The space \mathcal{L}' of all **linear functionals** on \mathcal{L} , i.e. mappings Λ from \mathcal{L} to \mathbb{R} such that $\Lambda(\lambda X + \mu Y) = \lambda\Lambda(X) + \mu\Lambda(Y)$ whenever $X, Y \in \mathcal{L}$ and $\lambda, \mu \in \mathbb{R}$, forms a linear space under the natural pointwise operations, with the zero mapping as its zero vector. By Corollary 2.8.5, the space \mathcal{P} of all linear previsions on \mathcal{L} is the subset of \mathcal{L}' containing all **positive** linear functionals with **unit norm**, i.e. all P in \mathcal{L}' such that $P(1) = 1$ and $P(X) \geq 0$ whenever $X \geq 0$.

D2 Linear topological spaces

A linear space \mathcal{X} is called a **linear topological space** (or **topological vector space**) when \mathcal{X} has a topology such that the two mappings $(x, y) \rightarrow x + y$ and $(\lambda, x) \rightarrow \lambda x$ are continuous under the product topologies on $\mathcal{X} \times \mathcal{X}$ and $\mathbb{R} \times \mathcal{X}$ respectively. We will assume that the topology on \mathcal{X} is **Hausdorff**, i.e. for any two distinct points x and y of \mathcal{X} , there are disjoint open sets \mathcal{V} and \mathcal{W} such that $x \in \mathcal{V}$, $y \in \mathcal{W}$. (See Kelley, 1955, for the basic ideas of topology.) Any linear space whose topology is generated by a norm is a linear topological space. Any linear topological space with finite dimension n is isomorphic to \mathbb{R}^n with its usual topology.

The class \mathcal{L} of all gambles is a **Banach space** (a complete normed space) under the **supremum norm** $\|X\| = \sup |X|$, and \mathcal{L} is a linear topological space under the topology generated by the supremum norm. The linear subspace of \mathcal{L} consisting of simple gambles (those taking only finitely many values) is a dense subset of \mathcal{L} in this topology.

A linear functional $\Lambda \in \mathcal{L}'$ is **continuous** if $\Lambda(X_n) \rightarrow 0$ whenever $\|X_n\| \rightarrow 0$. The space \mathcal{L}^* of all continuous linear functionals on \mathcal{L} is a linear subspace of \mathcal{L}' called the (topological or continuous) **dual space** of \mathcal{L} . All positive linear functionals are continuous. Thus the dual space \mathcal{L}^* contains the class \mathcal{P} of linear previsions.

D3 Weak* topology

This is defined as the weakest topology (i.e., the topology with the fewest open sets) on \mathcal{L}^* such that all the **evaluation functionals** X^* mapping \mathcal{L}^* to \mathbb{R} , defined by $X^*(\Lambda) = \Lambda(X)$ where $\Lambda \in \mathcal{L}^*$ and $X \in \mathcal{L}$, are continuous. (In fact, a linear functional on \mathcal{L}^* is weak*-continuous if and only if it is such an evaluation functional.) The weak* topology on the dual space \mathcal{L}^* induces a topology on the class \mathcal{P} of linear previsions by forming intersections with

\mathcal{P} of the open sets. The neighbourhoods of $P_0 \in \mathcal{P}$ under the weak* topology are the sets containing $\{P \in \mathcal{P} : |P(X_i) - P_0(X_i)| \leq 1 \text{ for } 1 \leq i \leq n\}$ for some $X_i \in \mathcal{L}$.

D4 Compactness of \mathcal{P}

The evaluation functionals X^* are weak*-continuous, and the half-spaces $\{\Lambda \in \mathcal{L}^* : \Lambda(X) \geq \lambda\}$ defined by $X \in \mathcal{L}$ and $\lambda \in \mathbb{R}$ are therefore weak*-closed. Hence the class of linear previsions $\mathcal{P} = \bigcap_{X \geq 0} \{\Lambda \in \mathcal{L}^* : \Lambda(X) \geq 0\} \cap \{\Lambda \in \mathcal{L}^* : \Lambda(1) = 1\}$ is a weak*-closed subset of \mathcal{L}^* .

A set \mathcal{T} is said to be **compact** if for every collection of open sets that covers \mathcal{T} there is a finite sub-collection that covers \mathcal{T} . Compact sets are closed, and closed subsets of compact sets are compact.

The Alaoglu–Bourbaki theorem (Holmes, 1975, p. 70), or Tychonoff's theorem (Kelley, 1955, Theorem 7.1), imply that \mathcal{P} is weak*-compact. It follows that the subsets \mathcal{M} of \mathcal{P} that are weak*-compact are just those that are weak*-closed. (However, \mathcal{P} is not sequentially compact under the weak* topology: not every sequence in \mathcal{P} has a weak*-convergent subsequence.)

The weak* topology really is very weak. That is indicated by the fact that the set of linear previsions P with **finite support**, for which there is a finite subset A of Ω such that $P(A) = 1$, is weak*-dense in \mathcal{P} . Thus \mathcal{P} is the smallest weak*-closed set containing all linear previsions with finite support, and \mathcal{P} is the smallest weak*-closed convex set containing all degenerate linear previsions.

It follows from this, when Ω is infinite, that the class \mathcal{P}_c of countably additive probabilities (defined on a σ -field of subsets of Ω that contains all singletons) is not weak*-closed. To see that, note that \mathcal{P}_c contains all finitely additive probabilities with finite support, so the weak*-closure of \mathcal{P}_c is \mathcal{P} . But \mathcal{P}_c is strictly contained in \mathcal{P} , since by Theorem 3.6.8 there are 0–1 valued additive probabilities such that $P(A) = 1$ for some countable set A but $P(B) = 0$ for all finite subsets B of A .

APPENDIX E

Separating hyperplane theorems

Versions of the separating hyperplane theorem were needed to prove the lower envelope theorem 3.3.3 and the weak*-compactness theorem 3.6.1, which were used to represent coherent lower previsions as lower envelopes of linear previsions. The topological version of the separating hyperplane theorem (E1 below) is equivalent to the Hahn–Banach theorem and to various other important results listed by Holmes (1975, pp. 24 and 95). Proofs of most of the results stated here can be found in Holmes' book and in many other texts. Although the separating hyperplane and Hahn–Banach theorems are usually proved using a version of the axiom of choice (e.g. Holmes uses Zorn's lemma), they are known to be strictly weaker than the ultrafilter theorem 3.6.5, which is in turn weaker than the axiom of choice. (See note 12 of section 3.6.)

A subset \mathcal{V} of a linear space \mathcal{Z} is said to be **convex** when $\lambda x + (1 - \lambda)y \in \mathcal{V}$ for all $x, y \in \mathcal{V}$ and $0 < \lambda < 1$. The **convex hull** of \mathcal{V} , written $\text{co}(\mathcal{V})$, is the intersection of all convex sets containing \mathcal{V} . Equivalently, $\text{co}(\mathcal{V}) = \{\sum_{i=1}^n \lambda_i v_i : n \geq 1, \lambda_i > 0, \sum_{i=1}^n \lambda_i = 1, v_i \in \mathcal{V}\}$.

E1 Separating hyperplane theorem (Holmes, p. 63)

Let \mathcal{V} and \mathcal{W} be convex subsets of a linear topological space \mathcal{Z} , and suppose \mathcal{V} has non-empty topological interior $\text{int}(\mathcal{V})$. There is a non-zero continuous linear functional Λ on \mathcal{Z} and real number α , such that $\Lambda(v) \geq \alpha$ for all $v \in \mathcal{V}$ and $\Lambda(w) \leq \alpha$ for all $w \in \mathcal{W}$, if and only if $\text{int}(\mathcal{V}) \cap \mathcal{W} = \emptyset$. If also $0 \in \mathcal{V} \cap \mathcal{W}$ then $\alpha = 0$.

The continuous linear functional Λ in the theorem defines a **closed hyperplane** $\mathcal{H} = \{x \in \mathcal{Z} : \Lambda(x) = \alpha\}$ which **separates** \mathcal{V} from \mathcal{W} . The sets \mathcal{V} and \mathcal{W} may be only weakly ‘separated’ by the hyperplane \mathcal{H} , in that substantial subsets of them may actually lie on \mathcal{H} . (Indeed, it is possible that \mathcal{H} contains both sets.) It follows from the theorem that \mathcal{V} and \mathcal{W} can be ‘strictly separated’ by a continuous linear functional Λ , with $\Lambda(v) > 0$ for all $v \in \mathcal{V}$ and $\Lambda(w) \leq 0$ for all $w \in \mathcal{W}$, if and only if there is an open convex set \mathcal{T} such that $\mathcal{T} \supset \mathcal{V}$, $\text{closure}(\mathcal{T}) \supset -\mathcal{W}$, and $0 \notin \mathcal{T}$. When \mathcal{Z} is

finite-dimensional, \mathcal{V} and \mathcal{W} are finite and \mathcal{V} is non-empty, this condition holds if and only if there are no $v_i \in \mathcal{V}, w_j \in \mathcal{W}, \lambda_i \geq 0$ with some $\lambda_i > 0$ and $\mu_j \geq 0$ such that $\sum_{i=1}^m \lambda_i v_i - \sum_{j=1}^n \mu_j w_j = 0$. (This is a version of the **theorem of the alternative**, e.g. Gale, 1960, Theorems 2.9 and 2.10.) Necessary and sufficient conditions for a stronger type of separation are given in Theorem E3.

The requirement in the theorem that one of the sets to be separated must have a non-empty interior can be removed when the linear space \mathcal{Z} is finite-dimensional. In that case \mathcal{Z} is isomorphic to \mathbb{R}^n , where n is the dimension of \mathcal{Z} , and all linear functionals on \mathcal{Z} are continuous.

E2 Theorem (Holmes, p. 15)

Any two disjoint convex subsets of a finite-dimensional linear topological space can be separated by a non-zero linear functional. (In fact, it is not necessary that the two sets be disjoint – it suffices that their relative interiors be disjoint.)

A linear topological space is called a **locally convex space** when every neighbourhood of zero contains a convex neighbourhood of zero. Every normed space is a locally convex space, since every neighbourhood of zero contains a convex neighbourhood $\{x : \|x\| < \delta\}$. The dual space \mathcal{L}^* is locally convex under the weak* topology.

In locally convex spaces, a strong kind of separation can be characterized as follows.

E3 Strong separation theorem (Holmes, p. 64)

Let \mathcal{V} and \mathcal{W} be convex subsets of a locally convex space \mathcal{Z} . There is a continuous linear functional Λ on \mathcal{Z} , $\alpha \in \mathbb{R}$ and $\delta > 0$, such that $\Lambda(v) \geq \alpha + \delta$ for all $v \in \mathcal{V}$ and $\Lambda(w) \leq \alpha - \delta$ for all $w \in \mathcal{W}$, if and only if 0 is not in the closure of $\{v - w : v \in \mathcal{V}, w \in \mathcal{W}\}$. Then the closed hyperplane $\mathcal{H} = \{x \in \mathcal{Z} : \Lambda(x) = \alpha\}$ is said to **strongly separate** \mathcal{V} and \mathcal{W} . In particular, \mathcal{V} and \mathcal{W} can be strongly separated if \mathcal{V} is closed convex, \mathcal{W} is compact convex, and $\mathcal{V} \cap \mathcal{W} = \emptyset$.

E4 Corollary

If \mathcal{V} is a closed convex subset of a locally convex space \mathcal{Z} , and x is in \mathcal{Z} but not in \mathcal{V} , then there is a continuous linear functional Λ on \mathcal{Z} and $\alpha \in \mathbb{R}$ such that $\Lambda(x) < \alpha$ and $\Lambda(v) \geq \alpha$ for all $v \in \mathcal{V}$.

Next we state a version of the Krein–Milman theorem, which shows that a compact convex set can be recovered as the closed convex hull of its set

of extreme points. Call v an **extreme point** of \mathcal{V} when v cannot be written as a convex combination of distinct elements of \mathcal{V} , i.e. when $v = \lambda x + (1 - \lambda)y$ for $x, y \in \mathcal{V}$ and $0 < \lambda < 1$ implies $x = y = v$. Let $\text{ext } \mathcal{V}$ denote the set of all extreme points of \mathcal{V} .

E5 Krein–Milman theorem (Holmes, p. 74)

Suppose \mathcal{Z} is a locally convex space and \mathcal{V} is a non-empty compact convex subset of \mathcal{Z} . Then

- (a) $\text{ext } \mathcal{V}$ is non-empty
- (b) \mathcal{V} is the closed convex hull of $\text{ext } \mathcal{V}$, i.e. the intersection of all closed convex sets containing $\text{ext } \mathcal{V}$, i.e. the closure of $\text{co}(\text{ext } \mathcal{V})$
- (c) every continuous linear functional on \mathcal{Z} attains its minimum on \mathcal{V} at an extreme point of \mathcal{V} .

When \mathcal{Z} is finite-dimensional, the operation of closure is not needed: any compact convex \mathcal{V} is the convex hull of its set of extreme points, $\mathcal{V} = \text{co}(\text{ext } \mathcal{V})$.

Finally, we state a general version of von Neumann's minimax theorem.

E6 Minimax theorem (Stoer and Witzgall, 1970, Theorem 6.3.7)

Suppose \mathcal{V} and \mathcal{W} are compact convex sets in linear topological spaces, and Π is a real function on $\mathcal{V} \times \mathcal{W}$ with the two properties

- (a) $\{v \in \mathcal{V} : \Pi(v, w) \leq \alpha\}$ is closed and convex for all $w \in \mathcal{W}$ and $\alpha \in \mathbb{R}$
- (b) $\{w \in \mathcal{W} : \Pi(v, w) \geq \alpha\}$ is closed and convex for all $v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$.

(These two properties hold, for example, if $\Pi(\cdot, w)$ is a continuous convex function for all $w \in \mathcal{W}$, and $\Pi(v, \cdot)$ is a continuous concave function for all $v \in \mathcal{V}$.) Then

$$\min_{v \in \mathcal{V}} \max_{w \in \mathcal{W}} \Pi(v, w) = \max_{w \in \mathcal{W}} \min_{v \in \mathcal{V}} \Pi(v, w).$$

Moreover, Π has a **saddle point** (v_0, w_0) such that

$$\Pi(v_0, w) \leq \Pi(v_0, w_0) \leq \Pi(v, w_0) \quad \text{for all } v \in \mathcal{V} \text{ and } w \in \mathcal{W},$$

and (v_0, w_0) attains both the min-max and the max-min in the above equality.

APPENDIX F

Desirability

There are strong reasons for taking a class \mathcal{R} of desirable gambles to be the fundamental model for uncertainty, as suggested in section 3.8.6. Many classes \mathcal{R} will induce the same lower prevision \underline{P} , and the extra information in \mathcal{R} is essential for defining non-vacuous previsions conditional on an event of probability zero. In this appendix we suggest appropriate coherence axioms for \mathcal{R} , and indicate how it determines coherent conditional previsions.

The behavioural interpretation is that You are disposed to accept each gamble in \mathcal{R} . Thus \mathcal{R} contains the ‘really desirable’ gambles. Compare with the classes \mathcal{D} and \mathcal{D}^+ of almost-desirable and strictly desirable gambles, considered in section 3.7. In general, $\mathcal{D} \supset \mathcal{R} \supset \mathcal{D}^+$. If the three classes are coherent then \mathcal{D} is the relative closure of \mathcal{R} , and \mathcal{D}^+ is its relative interior plus the non-negative non-zero gambles.

Coherence of \mathcal{R} is characterized by the following axioms. Here we require (in D11) that the zero gamble is in \mathcal{R} . This assumption slightly simplifies the axioms, especially the conglomerability axiom D12.

F1 Coherence axioms

Suppose that $\mathcal{L} \supset \mathcal{K} \supset \mathcal{R}$, where \mathcal{K} is a linear space containing constant gambles. Say that \mathcal{R} is **coherent relative to \mathcal{K}** if it satisfies the five axioms:

- (D2) if $X \in \mathcal{R}$ and $\lambda > 0$ then $\lambda X \in \mathcal{R}$ (positive homogeneity)
- (D3) if $X \in \mathcal{R}$ and $Y \in \mathcal{R}$ then $X + Y \in \mathcal{R}$ (addition)
- (D10) if $X \leq 0$ and $X \neq 0$ then $X \notin \mathcal{R}$ (avoiding partial loss)
- (D11) if $X \in \mathcal{K}$ and $X \geq 0$ then $X \in \mathcal{R}$ (accepting partial gains)
- (D12) if $X \in \mathcal{K}$, \mathcal{B} is a partition of Ω and $BX \in \mathcal{R}$ for all $B \in \mathcal{B}$, then $X \in \mathcal{R}$ (conglomerability).

The first four axioms, which were suggested by Williams (1975b), are the same as the first four axioms for strict desirability (3.7.8), except that we have now required the zero gamble to be desirable. The openness axiom D7 for strict desirability need not be satisfied by \mathcal{R} . The conglomerability

axiom D12 is the basis for our theory of conditioning and statistical inference. It was used to justify the coherence axioms for conditional previsions, in sections 6.3 and 7.1. When applied to a single partition \mathcal{B} , D12 corresponds to the \mathcal{B} -conglomerability axiom P7. The general version corresponds to the full conglomerability axiom P8 (see section 6.8).

F2 Induced conditional previsions

Suppose \mathcal{R} is coherent relative to \mathcal{K} , where \mathcal{K} is a linear space containing constant gambles. Then \mathcal{R} induces a coherent lower prevision \underline{P} on domain \mathcal{K} by $\underline{P}(X) = \sup\{\mu: X - \mu \in \mathcal{R}\}$. It also induces conditional lower previsions. Let \mathcal{B} be any partition of Ω such that $BX \in \mathcal{K}$ whenever $X \in \mathcal{K}$ and $B \in \mathcal{B}$. Define the conditional lower prevision $\underline{P}(\cdot|\mathcal{B})$ on \mathcal{K} by $\underline{P}(X|B) = \sup\{\mu: B(X - \mu) \in \mathcal{R}\}$. One way to see that this definition is reasonable is to consider $\mathcal{R}^B = \{Y \in \mathcal{K}: BY \in \mathcal{R}\}$, which is the class of gambles that are desirable after B has been observed, and note that $\underline{P}(X|B) = \sup\{\mu: X - \mu \in \mathcal{R}^B\}$ is the unconditional prevision induced by the updated class \mathcal{R}^B . Thus \mathcal{R} contains all the information needed to define conditional, as well as unconditional, previsions.

The first four coherence axioms are sufficient to ensure that $\underline{P}(\cdot|\mathcal{B})$ is separately coherent. When the conglomerability axiom D12 is added, they guarantee coherence in the much stronger sense of condition 7.1.4. (That can be proved by following the justification of conditions 7.1.2 and 7.1.4.)

F3 Theorem

Suppose that \mathcal{R} satisfies the coherence axioms in section F1. Let $\mathcal{B}_1, \dots, \mathcal{B}_n$ be partitions of Ω , and define each $\underline{P}(\cdot|\mathcal{B}_j)$ on domain \mathcal{K} by $\underline{P}(X|B) = \sup\{\mu: B(X - \mu) \in \mathcal{R}\}$ for $X \in \mathcal{K}$ and $B \in \mathcal{B}_j$. Then $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_n)$ are coherent, in the sense of 7.1.4(b).

F4 Conditioning on events of probability zero

Different classes \mathcal{R} that induce the same unconditional prevision \underline{P} can induce different conditional previsions. First consider

$$\mathcal{R}_1 = \{X \in \mathcal{L}: \underline{P}(X) > 0 \text{ or } X \geq 0\},$$

which is the class of strictly desirable gambles that corresponds to \underline{P} (as in 3.8.1) plus the zero gamble. Provided \underline{P} is a coherent lower prevision on domain \mathcal{L} and is fully conglomerable, \mathcal{R}_1 is coherent relative to \mathcal{L} . It induces the conditional previsions

$$\underline{E}(X|B) = \sup\{\mu: B(X - \mu) \in \mathcal{R}_1\} = \sup\{\mu: \underline{P}(B(X - \mu)) > 0 \text{ or } B(X - \mu) \geq 0\}.$$

This is just the natural extension of \underline{P} to a conditional prevision, defined in 6.8.2(c) and 8.1.6.

The larger class of desirable gambles

$$\mathcal{R}_2 = \{X \in \mathcal{L} : \underline{P}(X) \geq 0 \text{ and } \bar{P}(X) > 0, \text{ or } X \geq 0\},$$

which is coherent relative to \mathcal{L} provided \underline{P} satisfies the regularity axiom P9 (Appendix J3) for every partition, induces the conditional previsions

$$\begin{aligned}\underline{R}(X|B) &= \sup \{\mu : B(X - \mu) \in \mathcal{R}_2\} \\ &= \sup \{\mu : \underline{P}(B(X - \mu)) \geq 0 \text{ and } \bar{P}(B(X - \mu)) > 0, \text{ or } B(X - \mu) \geq 0\}.\end{aligned}$$

This is the regular extension of \underline{P} , defined in Appendix J. When $\underline{P}(B) = 0$ but $\bar{P}(B) > 0$, the regular extension may differ from the natural extension, which is vacuous.

The regular extension is also vacuous whenever $\bar{P}(B) = 0$, but then a different class \mathcal{R} can induce non-vacuous conditional previsions. For example, any separately-coherent conditional previsions $\underline{P}(\cdot|\mathcal{B})$ can be generated by

$$\mathcal{R}_3 = \{X \in \mathcal{L} : (\forall B \in \mathcal{B}) \underline{P}(X|B) > 0 \text{ or } BX \geq 0\}.$$

(This is coherent relative to \mathcal{L} provided each conditional prevision $\underline{P}(\cdot|B)$ is fully conglomerable.) If $\underline{P}(\cdot|\mathcal{B})$ is coherent with \underline{P} , both can be generated by the class $\mathcal{R}_4 = \mathcal{R}_1 \cup \mathcal{R}_3$. (This is coherent relative to \mathcal{L} provided \underline{P} and $\underline{P}(\cdot|B)$ are fully conglomerable.) Thus the extra information in the model \mathcal{R} can be useful for determining conditional previsions, even though it concerns only the marginally desirable gambles, those with lower prevision zero.

F5 Preferences

The class of desirable gambles \mathcal{R} corresponds to a partial preference ordering on gambles \geqslant , by $X \geqslant Y$ if and only if $X - Y \in \mathcal{R}$. The behavioural interpretation of $X \geqslant Y$ is that You are willing to give up Y in return for X . When \geqslant is defined on a linear space \mathcal{K} that contains constant gambles, coherence of \geqslant is characterized by the following axioms, which correspond to the axioms in F1.

- (R1) if $X, Y \in \mathcal{K}$ and $X \geq Y$ then $X \geq Y$ (monotonicity)
- (R2) if $X \geq Y$ and $\lambda > 0$ then $\lambda X \geq \lambda Y$ (positive homogeneity)
- (R3) if $X \geq Y$ and $Y \geq Z$ then $X \geq Z$ (transitivity)
- (R5) $X \geq Y$ if and only if $X - Y \geq 0$ (cancellation)
- (R11) if $X \geq Y$ and $X \neq Y$ then not $Y \geq X$ (strict monotonicity)
- (R12) if $X, Y \in \mathcal{K}$, \mathcal{B} is a partition of Ω and $BX \geq BY$ for all $B \in \mathcal{B}$, then $X \geq Y$ (conglomerability).

Upper and lower variances

The **variance** of a gamble X under a linear prevision P is defined as $V_P(X) = P((X - P(X))^2)$. This is one measure of the degree of uncertainty concerning X under P . (Specifically, it measures the expected squared distance of the reward $X(\omega)$ from its prevision $P(X)$.) In this appendix we show how the degree of uncertainty can be measured more generally, in cases of indeterminacy, by upper and lower variances.

Because of the identity $P((X - \mu)^2) = V_P(X) + (P(X) - \mu)^2$, an alternative expression for the variance is $V_P(X) = \min \{P((X - \mu)^2) : \mu \in \mathbb{R}\}$, and the minimum is achieved by $\mu = P(X)$. We will define upper and lower variances by analogy with this expression, in terms of upper and lower previsions. Throughout this appendix, we assume that \underline{P} is a coherent lower prevision defined on all gambles, and \bar{P} is its conjugate upper prevision.

G1 Definition

For each gamble X , define the **lower variance** of X under \underline{P} to be $\underline{V}(X) = \min \{\underline{P}((X - \mu)^2) : \mu \in \mathbb{R}\}$, and the **upper variance** of X under \bar{P} to be $\bar{V}(X) = \min \{\bar{P}((X - \mu)^2) : \mu \in \mathbb{R}\}$. (We show in Theorem G2 that the two minima are attained.) Define the **lower and upper standard deviations** of X , $\underline{\sigma}(X)$ and $\bar{\sigma}(X)$, to be the positive square roots of $\underline{V}(X)$ and $\bar{V}(X)$.

These definitions have a behavioural interpretation, in terms of Your attitudes to gambles of the form $(X - \mu)^2$. The lower variance $\underline{V}(X)$ is the supremum price You are willing to pay for each of the gambles $(X - \mu)^2$, as μ takes all real values. The upper variance $\bar{V}(X)$ is the infimum price for which You are willing to sell some gamble $(X - \mu)^2$.

When $\underline{P} = \bar{P} = P$ is a linear prevision, the upper and lower variances both agree with the variance $V_P(X)$. The following theorem shows that, in general, $\bar{V}(X)$ and $\underline{V}(X)$ are just the maximum and minimum of variances $V_P(X)$ under the linear previsions P in $\mathcal{M}(\underline{P})$. Because of this, the upper and lower variances also have sensitivity analysis interpretations: if the true linear prevision P is known only to dominate \underline{P} then the true variance $V_P(X)$ is known only to lie between the upper and lower bounds $\bar{V}(X)$ and $\underline{V}(X)$.

G2 Variance envelope theorem

Let $\mathcal{M} = \mathcal{M}(\underline{P})$ denote the class of dominating linear previsions for \underline{P} , and \underline{V}, \bar{V} its lower and upper variances. Then $\underline{V}(X) = \min\{V_P(X): P \in \mathcal{M}\}$ and $\bar{V}(X) = \max\{V_P(X): P \in \mathcal{M}\}$ for all gambles X . The two minima in the definitions of $\underline{V}(X)$ and $\bar{V}(X)$ are attained by $\mu_1 = P_1(X)$ and $\mu_2 = P_2(X)$, where P_1 and P_2 respectively minimize and maximize $V_P(X)$ over P in \mathcal{M} .

Proof. Consider a fixed gamble X . Define functionals Π and Υ for $\mu \in \mathbb{R}$ and $P \in \mathcal{P}$ by

$$\Pi(\mu, P) = P((X - \mu)^2)$$

$$\text{and } \Upsilon(P) = V_P(X) = P(X^2) - (P(X))^2 = \min_{\mu \in \mathbb{R}} \Pi(\mu, P).$$

Then Υ is weak*-continuous on \mathcal{P} . Since \mathcal{M} is weak*-compact (Theorem 3.6.1), Υ attains its maximum and minimum over \mathcal{M} . Let $P_1 \in \mathcal{M}$ attain the minimum. Then

$$\begin{aligned} \Upsilon(P_1) &= \min_{P \in \mathcal{M}} \Upsilon(P) = \min_{P \in \mathcal{M}} \min_{\mu \in \mathbb{R}} \Pi(\mu, P) = \min_{\mu \in \mathbb{R}} \min_{P \in \mathcal{M}} \Pi(\mu, P) \\ &= \min_{\mu \in \mathbb{R}} \underline{P}((X - \mu)^2) = \underline{V}(X), \end{aligned}$$

and the minima are attained by $P = P_1$ and $\mu = P_1(X)$.

Verify that $\bar{P}((X - \mu)^2) > \bar{P}((X - \underline{P}(X))^2)$ when $\mu < \underline{P}(X)$, and $\bar{P}((X - \mu)^2) > \bar{P}((X - \bar{P}(X))^2)$ when $\mu > \bar{P}(X)$. To minimize $\bar{P}((X - \mu)^2)$, we can therefore restrict attention to values of μ in the closed interval $\mathcal{V} = [\underline{P}(X), \bar{P}(X)]$. Since Π is a linear function of P and a convex (quadratic) function of μ , the minimax theorem (E6) can be applied to show

$$\begin{aligned} \Upsilon(P_2) &= \max_{P \in \mathcal{M}} \Upsilon(P) = \max_{P \in \mathcal{M}} \min_{\mu \in \mathcal{V}} \Pi(\mu, P) = \min_{\mu \in \mathcal{V}} \max_{P \in \mathcal{M}} \Pi(\mu, P) \\ &= \min_{\mu \in \mathcal{V}} \bar{P}((X - \mu)^2) = \bar{V}(X), \end{aligned}$$

where P_2 maximizes Υ over \mathcal{M} . Let μ_2 minimize $\bar{P}((X - \mu)^2)$. Then

$$\begin{aligned} \bar{V}(X) &= \bar{P}((X - \mu_2)^2) \geq P_2((X - \mu_2)^2) = \Pi(\mu_2, P_2) \geq \Pi(P_2(X), P_2) \\ &= \Upsilon(P_2) = \bar{V}(X), \end{aligned}$$

so there is equality everywhere. Hence $\mu_2 = P_2(X)$. \blacklozenge

G3 Properties of P_1 and P_2

By concavity of the variance functional Υ , the minimizing value $\mu_2 = P_2(X)$ in the definition of $\bar{V}(X)$ is uniquely determined, whereas $\mu_1 = P_1(X)$ is

typically not unique. In section G6, for example, μ_1 takes all values in the interval $[\underline{P}(X), \bar{P}(X)]$.

Both μ_1 and μ_2 can be regarded as point estimates of the uncertain quantity X under the quadratic loss function, where Your loss from estimate μ when ω is the true state is $(X(\omega) - \mu)^2$. In fact, μ_2 is the unique \underline{P} -minimax estimate of X (in the sense of section 3.9.7), since it minimizes $\bar{P}((X - \mu)^2)$, the upper revision of the loss. The minimized value is just the upper variance $\bar{V}(X)$. (Also μ_2 is the Bayes estimate under P_2 , the least favourable prevision in \mathcal{M} .) Similarly μ_1 is a (not necessarily unique) \underline{P} -minimin estimate, which minimizes the lower revision of the loss, and the minimized value is $\underline{V}(X)$. Other estimates are reasonable here. By Theorem 3.9.5, all the points in the interval $[\underline{P}(X), \bar{P}(X)]$ are maximal estimates of X under quadratic loss.

It is usually quite easy to calculate $\underline{V}(X)$ because P_1 can always be taken to be an extreme point of \mathcal{M} (again by concavity of Υ). Usually P_2 is not an extreme point (see examples G6 and G7), and the calculation of $\bar{V}(X)$ is more difficult. One useful method is to find a linear prevision P_2 such that $P((X - P_2(X))^2)$ is maximized over \mathcal{M} by $P = P_2$. Writing $\mu_2 = P_2(X)$, it follows that (μ_2, P_2) is a saddle point of Π , meaning that $P((X - \mu_2)^2) \leq \bar{P}((X - \mu_2)^2) = P_2((X - \mu_2)^2) \leq P_2((X - \mu)^2)$ for all real μ and $P \in \mathcal{M}$, and hence that the upper variance $\bar{V}(X)$ is just the variance under P_2 . (The pair (μ_2, P_2) constructed in the proof of Theorem G2 is such a saddle point, by the last step of the proof.) This method is illustrated in example G7.

G4 Basic inequalities

It can be verified that the upper and lower variances satisfy the following inequalities, in which we use the notation $\underline{\mu} = \underline{P}(X)$, $\bar{\mu} = \bar{P}(X)$, $\rho = \frac{1}{2}\underline{P}(X) + \frac{1}{2}\bar{P}(X)$ and $\Delta = \Delta(X) = \bar{P}(X) - \underline{P}(X)$.

- (1) $0 \leq \underline{V}(X) \leq \bar{V}(X)$
- (2) $\underline{V}(X) \leq \underline{P}((X - \underline{\mu})^2) \leq \underline{V}(X) + \Delta^2$, $\bar{V}(X) \leq \bar{P}((X - \bar{\mu})^2) \leq \bar{V}(X) + \Delta^2$
(and similarly when μ is replaced by $\bar{\mu}$)
- (3) $\underline{V}(X) \leq \underline{P}(X^2) \leq \underline{V}(X) + \max\{\mu^2, \bar{\mu}^2\}$
 $\bar{V}(X) \leq \bar{P}(X^2) \leq \bar{V}(X) + \max\{\underline{\mu}^2, \bar{\mu}^2\}$
- (4) $\underline{V}(X) \leq \underline{P}((X - \rho)^2) \leq \underline{V}(X) + \frac{1}{4}\Delta^2$
 $\bar{V}(X) \leq \bar{P}((X - \rho)^2) \leq \bar{V}(X) + \frac{1}{4}\Delta^2$
- (5) $\frac{1}{4}\Delta^2 \leq \bar{V}(X) - \underline{V}(X) \leq \frac{1}{4}\Delta^2 + \bar{P}((X - \rho)^2) - \underline{P}((X - \rho)^2)$
Hence $\bar{V}(X) \geq \frac{1}{4}\Delta^2$ and $\bar{\sigma}(X) \geq \frac{1}{2}\Delta$.

From (5), it is not possible for the upper and lower variances of X to be equal unless its upper and lower previsions are equal. This is essentially because the variance functional Υ is concave on \mathcal{M} ; the variance of X under

$\frac{1}{2}(P+Q)$ is strictly larger than $\frac{1}{2}V_P(X) + \frac{1}{2}V_Q(X)$ unless $P(X) = Q(X)$. In fact, a necessary and sufficient condition for $\underline{V}(X) = \bar{V}(X)$ is that X has the same prevision and the same variance under all extreme points of $\mathcal{M}(P)$.

To show how upper and lower variances can be calculated, we give three examples.

G5 Constant mean

Suppose that $\underline{P}(X) = \bar{P}(X) = \mu$, so all previsions in \mathcal{M} assign X the same mean μ . Then

$$\begin{aligned}\underline{V}(X) &= \underline{P}((X - \mu)^2) = \min\{V_P(X) : P \in \text{ext } \mathcal{M}\}, \\ \bar{V}(X) &= \bar{P}((X - \mu)^2) = \max\{V_P(X) : P \in \text{ext } \mathcal{M}\},\end{aligned}$$

and most of the inequalities in section G4 become equalities.

Consider the model of section 5.4.1, for example, where \underline{P} is the lower envelope of beta (s, t) distributions with fixed mean t and $s < s < \bar{s}$. Since the identity gamble $I(\theta) = \theta$ has variance $t(1-t)/(s+1)$ under the beta (s, t) distribution, we obtain upper and lower variances $\bar{V}(I) = t(1-t)/(s+1)$ and $\underline{V}(I) = t(1-t)/(\bar{s}+1)$.

G6 Constant variance

Suppose next that all extreme points of $\mathcal{M}(P)$ assign the same variance η^2 to X . Then $\underline{V}(X) = \eta^2$, and it can be seen from (4) and (5) of section G4 that $\bar{V}(X) = \eta^2 + \frac{1}{4}\Delta(X)^2$. The saddle-point prevision P_2 is a convex combination of previsions which attain $\underline{P}(X)$ and $\bar{P}(X)$, with $P_2(X) = \frac{1}{2}\underline{P}(X) + \frac{1}{2}\bar{P}(X)$.

For example, let the extreme points of \mathcal{M} be the Normal (μ, η^2) priors for θ , with constant variance η^2 and $\mu \leq \mu \leq \bar{\mu}$. Then the gamble $I(\theta) = \theta$ has $\underline{V}(I) = \eta^2$ and $\bar{V}(I) = \eta^2 + \frac{1}{4}(\bar{\mu} - \mu)^2$.

This example can be developed to give some insight into the large-sample behaviour of upper and lower standard deviations. Suppose we obtain n independent observations from a Normal (θ, σ^2) distribution with known σ^2 . The prior class \mathcal{M} for the unknown mean θ produces a posterior class \mathcal{M}_n of the same form, with η^2 , $\underline{\mu}$ and $\bar{\mu}$ updated to $\eta_n^2 = (\eta^{-2} + n\sigma^{-2})^{-1}$, $\underline{\mu}_n = \eta_n^2(\eta^{-2}\underline{\mu} + n\sigma^{-2}\bar{x})$ and $\bar{\mu}_n = \eta_n^2(\eta^{-2}\bar{\mu} + n\sigma^{-2}\bar{x})$, where \bar{x} is the sample mean. The dimensionless quantity $c = \sigma\eta^{-2}(\bar{\mu} - \mu)$ is constant under sampling. The posterior imprecision concerning \bar{I} is $\Delta_n(I) = \bar{\mu}_n - \mu_n = \eta_n^2\sigma^{-1}c$, and the posterior lower and upper standard deviations are $\underline{\sigma}_n(\bar{I}) = \eta_n$, $\bar{\sigma}_n(\bar{I}) = \eta_n(1 + \frac{1}{4}\eta_n^2\sigma^{-2}c^2)^{1/2}$. As $n \rightarrow \infty$, we obtain the approximations $\Delta_n(I) \simeq c\sigma n^{-1}$, $\underline{\sigma}_n(I) \simeq \sigma n^{-1/2}$ and $\bar{\sigma}_n(I) - \underline{\sigma}_n(I) \simeq \frac{1}{8}c^2\sigma n^{-3/2}$. Thus $\underline{\sigma}_n(I)$ and $\bar{\sigma}_n(I)$ are of order $n^{-1/2}$, $\Delta_n(I)$ is of order n^{-1} , and $\bar{\sigma}_n(I) - \underline{\sigma}_n(I)$ is of order $n^{-3/2}$. In particular, the degree of imprecision $\Delta_n(I)$ tends to zero

faster than the upper and lower standard deviations. A wide class of statistical models exhibit the same asymptotic behaviour.

G7 Beta–Bernoulli model

As a more substantial example, where neither the mean nor the variance is constant over extreme points of \mathcal{M} , consider the beta–Bernoulli model (section 5.3.6). Here \underline{P} is the lower envelope of beta (s, t) priors for the chance θ , with fixed s and $\underline{t} < t < \bar{t}$. This can be regarded as the posterior for a near-ignorance (S) prior after observing m successes in n Bernoulli trials, where $s = S + n$, $\underline{t} = m/(S+n)$ and $\bar{t} = (S+m)/(S+n)$. We compute the upper and lower variances of the identity gamble $I(\theta) = \theta$.

Let $V_t(I) = t(1-t)/(s+1)$ denote the variance of I under beta (s, t) . We obtain the lower variance simply by minimizing $V_t(I)$ over $\underline{t} \leq t \leq \bar{t}$. Hence

$$\underline{V}(I) = \min\{\underline{t}(1-\underline{t}), \bar{t}(1-\bar{t})\}/(s+1) = \frac{m(n-m) + S \min\{m, n-m\}}{(S+n)^2(S+n+1)}.$$

To obtain $\bar{V}(I)$ it is enough to find a saddle-point distribution P_2 such that $P = P_2$ maximizes $P((I - P_2(I))^2)$ over \mathcal{M} . Defining $\mu = \frac{1}{2}(s(\underline{t} + \bar{t}) + 1)/(s+1) = \frac{1}{2}(S+2m+1)/(S+n+1)$ and assuming that $S \geq 1$ to ensure that $\underline{t} < \mu < \bar{t}$, it can be verified that the maximum of $P((I - \mu)^2)$ when P is a beta (s, t) distribution with $\underline{t} \leq t \leq \bar{t}$ is attained by both $t = \underline{t}$ and $t = \bar{t}$. It follows that a suitable saddle point P_2 is the convex combination of beta (s, \underline{t}) and beta (s, \bar{t}) distributions with weights proportional to $\bar{t} - \mu$ and $\mu - \underline{t}$ respectively, which has mean $P_2(I) = \mu$ and variance $P_2((I - \mu)^2) = \mu^2 - \bar{t}\underline{t}/(s+1) = \mu^2 - m(S+m)/(S+n)(S+n+1) = \bar{V}(I)$, the upper variance.

The dependence of the posterior upper and lower standard deviations $\bar{\sigma}_n(I)$ and $\underline{\sigma}_n(I)$ on the sample size n is illustrated in the next table, based on the near-ignorance prior (section 5.3.2) with $S = 2$ and observed relative frequency $m/n = \frac{1}{2}$ (so that $\mu = \frac{1}{2}$).

n	0	2	4	10	20	100	1000
$\bar{\sigma}_n(I)$	0.5	0.32	0.24	0.16	0.113	0.0502	0.01582
$\underline{\sigma}_n(I)$	0	0.19	0.18	0.14	0.104	0.0493	0.01579
$\Delta_n(I)$	1	0.50	0.33	0.16	0.10	0.020	0.0020

The behaviour of $\bar{\sigma}_n(I)$, $\underline{\sigma}_n(I)$ and $\Delta_n(I)$ for large n is similar to that in example G6. Again, the degree of imprecision tends to zero faster than the upper and lower standard deviations as $n \rightarrow \infty$, although the three quantities are similar in magnitude for moderately large sample sizes.

APPENDIX H

Operational measurement procedures

The elicitation procedure discussed in Chapter 4 can be ‘operationalized’ by actually carrying out some of the gambles that You judge desirable. Some ways of doing this, involving different ways of selecting a subset of desirable gambles, will be described in this appendix. The aim is to reward the judgements You make during elicitation, in such a way that You have an incentive to report Your beliefs honestly. We pay special attention to the biases in each operational procedure that might encourage dishonesty. Throughout, we assume that all gambles are in units of probability currency (section 2.2), which acts as an interpersonal utility scale and preserves the linearity of utilities when many gambles are exchanged simultaneously.

H1 Advantages of operational measurement

The operational procedures described here should be compared with the general elicitation procedure of section 4.1. In the latter, Your probabilities are elicited through assertions that You are willing to accept specified gambles. In the operational procedures, You are required to accept some of these gambles. You must be prepared to back up Your assertions with actions, and put Your money where Your mouth is.

One reason for developing operational procedures is that assertions about behavioural dispositions are not always reliable. You may not take elicitation seriously enough to check that Your assertions reflect Your real beliefs. The assertions may be biased by conscious self-interest or by unconscious self-deception.

An expert on earthquake risk may think it highly unlikely that a large earthquake will occur within 50 miles of San Francisco in the next ten years, but report a probability somewhat higher than his real one because he fears that his assessment will be remembered only if such an earthquake occurs. Non-experts, such as residents of San Francisco or companies deciding whether to build a factory in the area, would probably prefer to know his real beliefs. In such cases it is useful to have practical elicitation procedures that encourage the expert to give an honest assessment.

It is not clear, however, that operational procedures will result in more accurate assessments of probabilities. That question requires investigation by experimental psychologists. For some discussion, see Hogarth (1975), Leamer (1986), Savage (1971), Winkler (1967b, 1971), von Winterfeldt and Edwards (1986, Ch. 11).

A general difficulty for all methods of elicitation is that You may have some personal interest in the state ω ; some states may have greater value for You than others. In such cases Your values, as well as Your beliefs about ω , may influence Your attitude to gambles. If, for example, the occurrence of some state ω_0 (such as the earthquake in San Francisco) would have unpleasant consequences for You, You may be tempted to insure Yourself against it by accepting gambles which are positive when ω_0 occurs, thereby distorting Your probability for ω_0 . The ‘three prisoners’ problem 5.13.10, in which ω_0 denotes execution rather than reprieve, is an extreme example of this.

These biases occur in both expert assessment and personal decision making. In decision making, the need to choose a rational decision already provides some incentive to disentangle beliefs from values and to assess probabilities accurately. It may be easier to do this if gambling is merely hypothetical, as in the elicitation procedure of Chapter 4.

Operational procedures seem likely to be more useful in eliciting probabilities from experts, such as weather forecasters or economic forecasters, who have no responsibility for making decisions. These experts are motivated by complex personal interests such as making money, preserving or enhancing their professional reputation, pleasing their clients, maintaining self-respect or advancing a cause. The rewards paid during elicitation, if they are sufficiently large, may help to overcome the biases produced by these personal interests.

There are several reasons for doubting whether operational procedures will be widely used. The procedures may be impracticable because of the time and money needed to apply them. The rewards from gambling will need to be quite large to overcome biases in elicitation, and You will be willing to participate only if Your expected reward is positive, so the procedures may be expensive to the elicitor. The procedures may fail to encourage honesty because they legitimize dishonesty. Your primary objective in gambling is to maximize Your personal profit, rather than to be honest. (In other situations, honesty is a strong moral obligation.) Finally, we will see that Your assessments may be distorted, in these operational procedures, by Your beliefs about which of the desirable gambles You will be given. This will tend to focus Your attention on game-theoretic aspects of elicitation, such as what the elicitor believes about ω , rather than on Your own beliefs.

Some operational procedures are useful for evaluating and comparing probability assessors. That can be done by arranging mutually acceptable bets between the assessors and comparing overall gain, or by directly comparing their 'scores'. An experiment that compared probability assessors in this way is described in Appendix I. Comparison of Your reward with that of other assessors, or with the reward You could have obtained from different assessments, may help You to discover biases and improve Your future assessments.

H2 Two-sided betting procedure

De Finetti (1931, 1974) and Mellor (1971, Ch. 2) have proposed operational definitions of probability that, in effect, require You to act as a bookmaker who must set two-sided betting rates. For a given gamble X You must specify some real number x , and You are thereby committed to accepting any gamble $\lambda(X - x)$. The stake λ , which may be positive or negative, will be chosen by me (the elicit) in the light of x . (We assume that $|\lambda|$ is bounded so that Your maximum possible loss is smaller than Your current fortune.) I will then pay You reward $\lambda(X(\omega) - x)$. The specified x measures Your prevision $P(X)$, which is assumed to be precise.

De Finetti (1974, section 3.2) argues that an operational definition is needed to ensure that probabilities have 'an effective meaning'. In our view, any operational definition of probability would unduly restrict the applicability of the concept. (Other arguments are outlined in section 2.10.2.) We will not consider the two-sided betting procedure as a definition of probability, but only as an elicitation technique that may be more or less accurate than others, as suggested by Borel (1943, section 3.8). Indeed there are several reasons why the two-sided betting procedure may measure probabilities inaccurately.

(a) Observability

It is assumed that the state ω will become known so that the gamble can be settled. (The procedure is practicable only if ω becomes known fairly quickly.) This is a difficulty with all procedures which pay rewards that are functions of ω . For alternative measurement procedures which avoid it, see section 2.11.9.

(b) Personal interest in ω

As noted in section H1, different states ω may have different values for You, in addition to the value of the gamble You receive, and this may distort Your assessment x . Again, this is a problem for all elicitation procedures.

(c) Indeterminacy

If Your beliefs are really indeterminate, the procedure does not allow You

to report them honestly even if You wish to do so. It is useful only if beliefs are determinate.

(d) Unfairness

The procedure is asymmetric between You and me, and clearly unfair to You. Suppose that You honestly report the precise prevision $x = P(X)$. If my own prevision $Q(X)$ is precise then I will choose positive stake λ if $Q(X) < P(X)$, and negative stake if $Q(X) > P(X)$. Even if my choice of stake provides no further information to You about ω , Your expected gain from the gamble will be zero whereas mine will be positive (unless $Q(X) = P(X)$). But typically You will 'respect' my beliefs in the sense that a choice of positive λ would lead You to reduce Your own $P(X)$, and negative λ to increase $P(X)$. It is then easy to see that Your expected gain from the procedure will be negative, whereas mine (if I modify my beliefs similarly in the light of x) will still be positive. Because the procedure is unfair, You may be unwilling to participate.

(e) Anticipating opponent's choice

If You have beliefs about my choice of stake λ , e.g. based on beliefs about my assessment $Q(X)$, it is advantageous for You to use these in choosing x different from $P(X)$. Typically You will choose x between $P(X)$ and Your prevision for $Q(X)$ in order to obtain a more favourable gamble. If You have indeterminate beliefs about $Q(X)$, You might choose x between $P(X)$ and $0.5(\inf X + \sup X)$ in order to reduce Your maximum possible loss. This problem also applies to the following procedures. The general problem is that Your gain from elicitation depends on choices made by the elicit, and Your beliefs about these may influence Your own choices.

H3 Bookmaking procedure

The problems of indeterminacy and unfairness can be resolved by allowing You to choose two real numbers x and y , thereby committing Yourself to accept the gambles $\lambda_1(X - x)$ and $\lambda_2(y - X)$ for any non-negative stakes λ_j chosen by me. (Again, λ_j must be smaller than some specified upper bound.) Then x is taken to measure Your lower prevision $\underline{P}(X)$ and y to measure Your upper prevision $\bar{P}(X)$. This procedure was suggested by Borel (1943, section 3.8). We call it the **bookmaking procedure** because bookmakers act in a similar way, setting different rates for bets on or against an event.

The bookmaking procedure is not unfair to You, since You can commit Yourself only to gambles that You would regard as favourable if You knew that Your opponent was willing to offer them. In the extreme case You can

effectively refuse to gamble by specifying the vacuous previsions $x = \inf X$ and $y = \sup X$.

The procedure can be extended by requiring You to specify a class of gambles \mathcal{D} . You will then be given an arbitrary finite combination of positive multiples of gambles in \mathcal{D} (with a suitable bound on the total stakes involved). This encourages \mathcal{D} to avoid sure loss (otherwise, You can be forced to incur sure loss).

Difficulty (e), anticipation of opponent's choice, remains and takes several forms which we will call anticipated information, low demand and low profitability. My choice of gambles from those You have offered may provide extra information which leads You to revise Your prices x and y . You can anticipate this information by choosing x so that $X - x$ is desirable conditional on my willingness to accept $x - X$. If You respect my beliefs, Your elicited intervals $[x, y]$ will be wider than Your true intervals $[\underline{P}(X), \bar{P}(X)]$ to reflect the anticipated information. The extent to which You widen Your intervals will depend on Your judgements about our relative amounts of information and expertise. This might be useful for some purposes such as aggregation of expert opinions. Then a lower rating of Your own expertise, relative to other assessors in a group, will lead to wider intervals which will have less influence on the aggregated beliefs of the group.

It may seem that, when You have less information than Your opponent, You should set vacuous intervals to avoid betting altogether. But that is overly cautious. If the opponent has bounded information then anticipating it should have bounded effect on Your probabilities. To illustrate that, suppose that You must assess probabilities $\underline{P}(A)$ and $\bar{P}(A)$ concerning the outcome of a Bernoulli trial. Your only relevant information is that You have observed m successes in n independent trials. Your opponent sees the same observations, but he also sees an extra n' independent observations. Suppose You adopt a near-ignorance (s) prior (5.3.2). Your posterior probabilities for A are then $\underline{P}_n(A) = m/(s+n)$ and $\bar{P}_n(A) = (s+m)/(s+n)$. If You saw the extra n' observations, Your lower probability for A would be at least $\underline{P}'(A) = m/(s+n+n')$, and Your upper probability would be no more than $\bar{P}'(A) = (s+m+n')/(s+n+n')$. (The difference between these bounds is an increasing function of $(s+n')/n$, and reflects the difference in information.)

Even if You are maximally cautious, and specify the bounds as Your elicited probabilities, You will be prepared to bet against a better-informed opponent at sufficiently attractive odds. Of course, if the opponent uses a similar model then the two probability intervals will intersect and no bets will be made. Betting is possible if he uses a different model (e.g. he has prior prejudices or a different way of processing the information), or if his

observations are independent of Yours but he does not widen his intervals to anticipate Your information.

H4 Demand and profitability

Anticipation of the opponent's choice can affect Your response $[x, y]$ in other ways, even when knowledge of Your opponent's beliefs would have no effect on Your previsions $\underline{P}(X)$ and $\bar{P}(X)$. These effects arise from Your desire to maximize Your expected profit, which depends on both the expected total stake (or **demand**) and the expected profit per unit stake (or **profitability**). Increasing the imprecision $y - x$ will tend to increase profitability but decrease demand. It follows that it will often be sensible for the elicited intervals $[x, y]$ to be wider than the true intervals $[\underline{P}(X), \bar{P}(X)]$ in cases of **low profitability**, when the true intervals are narrow, and for the elicited intervals to be narrower than the true intervals in cases of **low demand**, when the true intervals are wide.

Some insight into these effects can be gained by considering how bookmakers set their buying and selling prices for a gamble X . (Usually X is the indicator function of an event.) Suppose the bookmaker can assess precise prevision $\alpha_1(x)$ for the stakes bet against X at price x , i.e. the total number of gambles X that would be sold to him if his buying price was x . We assume that, conditional on x , these stakes are independent of y and ω . Similarly he assesses precise prevision $\alpha_2(y)$ for the stakes bet on X at price y , i.e. the total number of gambles X that would be bought from him if his selling price was y . (The same model applies to elicitation using the bookmaking procedure, with $\alpha_1(x)$ and $\alpha_2(y)$ Your previsions of my stake on or against X at various prices x and y , although the assumption of precision is then less realistic.) Assume that α_1 and α_2 are differentiable with derivatives β_1 and β_2 respectively. It is reasonable to suppose that α_1 is an increasing function and α_2 is decreasing, so that $\beta_1(x) \geq 0$ and $\beta_2(y) \leq 0$ for all real x, y . Conditional on his prices (x, y) and the outcome ω , the bookmaker has precise prevision $\rho(x, y, \omega) = \alpha_1(x)(X(\omega) - x) + \alpha_2(y)(y - X(\omega))$ for his overall profit.

Suppose first that the bookmaker attempts to make a book by equating the prevision of total stakes bet on X with those against X , so $\alpha_1(x) = \alpha_2(y)$. Then his prevision of overall profit is $\rho(x, y, \omega) = \alpha_1(x)(y - x)$, which does not depend on ω . He will choose (x, y) to maximize this, subject to the constraint $\alpha_1(x) = \alpha_2(y)$. So his expected profit is simply the product of expected demand for gambles, $\alpha_1(x)$, which increases with x , and the profit margin $y - x$, which decreases as x increases. Maximizing expected profit involves a trade-off between the profit margin and the expected demand.

Under simple regularity conditions, $\rho(x, y)$ can be maximized by equating its derivative to zero, giving the equations $y - x = \alpha_1(x)/\beta_1(x) - \alpha_2(y)/\beta_2(y)$ and $\alpha_1(x) = \alpha_2(y)$ for the optimal (x, y) .

Suppose, for example, that $X = A$ is an event, and Your beliefs about my precise probability $Q(A)$ are modelled by a uniform distribution on $[0, 1]$. Assume that, in the bookmaking procedure, I will choose stakes with upper bound λ to maximize my own expected profit, so I will stake λ against A at price x if and only if $Q(A) < x$, and I will stake λ on A at price y if and only if $Q(A) > y$. Then Your previsions for my stakes are $\alpha_1(x) = \lambda x$ and $\alpha_2(y) = \lambda(1 - y)$, with derivatives $\beta_1(x) = -\beta_2(y) = \lambda$. On the above model, the values x and y that maximize Your expected profit are given by $\lambda x = \lambda(1 - y)$ and $y - x = \frac{1}{2}$. Hence $x = \frac{1}{4}$ and $y = \frac{3}{4}$, with expected profit $\alpha_1(x)(y - x) = \lambda/8$.

Under this model, the bookmaker's odds do not depend in any way on his own beliefs about Ω , but only on his beliefs about the demand for bets. Even if he has vacuous beliefs about Ω , he will quote non-vacuous odds. He will try to accept equal numbers of gambles $X - x$ and $y - X$, even though neither is desirable for him in isolation, because together they give him a sure gain of $y - x$. He will not set $y = x$, even if he has precise prevision $P(X) = \bar{P}(X)$, because that would produce zero expected profit.

A Bayesian with precise prevision $P(X)$, rather than trying to make a book on X , would choose (x, y) to maximize the prevision of his profit, which is simply $\alpha_1(x)(P(X) - x) + \alpha_2(y)(y - P(X))$. By equating the partial derivatives to zero we obtain the equations $\beta_1(x)(P(X) - x) = \alpha_1(x)$ for x , and $-\beta_2(y)(y - P(X)) = \alpha_2(y)$ for y . Because α_1 is increasing and α_2 decreasing we have (except in degenerate cases) $x < P(X) < y$. The elicited interval width satisfies $y - x = \alpha_1(x)/\beta_1(x) - \alpha_2(y)/\beta_2(y)$, irrespective of beliefs about Ω , as for the previous model. In the previous numerical example we find $x = P(A)/2$, $y = \frac{1}{2} + P(A)/2$, and $y - x = \frac{1}{2}$ irrespective of $P(A)$. Thus, even on a standard Bayesian analysis, You should report imprecise previsions $x < P(X) < y$ when Your true prevision is precise and You have precise beliefs about which gambles will be accepted by the elicitor.

In practical elicitation it is likely that, contrary to the above models, Your beliefs about which gambles will be chosen by the elicitor will be much less precise than Your beliefs about Ω . Their effect on Your response is then harder to predict, but seems likely to be reduced.

H5 Multiple price procedure

The low profitability effect arises in the bookmaking procedure because transactions can take place only at the prices x and y that You set, and not at other prices that are more favourable to You but may also be accept-

able to me. This effect can be eliminated by allowing transactions to take place at a range of different prices.

Suppose that You specify two real numbers x and y as Your lower and upper previsions for X . In the light of Your choices, I specify numbers u and v . We then buy and sell X from each other at all the mutually acceptable prices z , such that $v < z < x$ or $y < z < u$, with stakes $\zeta(z)dz$. Here ζ is a pre-specified, positive function on the real line called the **stakes function**. Let $\chi(v, x)$ take the value one if $v < x$, and zero otherwise. Then Your net gain from the transactions involving X is the gamble

$$Z(u, v, x, y) = \chi(v, x) \int_v^x (X - z)\zeta(z)dz + \chi(y, u) \int_y^u (z - X)\zeta(z)dz.$$

This procedure can be used to simultaneously elicit Your upper and lower previsions for finitely many gambles.

My choices u and v might represent my own lower and upper previsions for X , but that is not essential. Since they may depend on x and y , the procedure is not symmetric between us. In particular, if Your assessments incurred sure loss then I could choose u and v to exploit that. However, this procedure, like the bookmaking procedure, is not unfair to You. The main advantage over the bookmaking procedure is that, provided You do not try to influence my choices, the low profitability effect is eliminated because it is in Your interest to allow transactions at all desirable prices.

H6 Symmetric procedure

The multiple price procedure can be easily modified to make it symmetric, by requiring both pairs (x, y) and (u, v) to be chosen in ignorance of the other's choice. More generally, we independently specify our classes of desirable gambles, \mathcal{D}_1 and \mathcal{D}_2 , and we then exchange all the mutually acceptable gambles, those in $\mathcal{D}_1 \cap (-\mathcal{D}_2)$, weighted according to some pre-specified stakes function.

This symmetric procedure is again fair to both participants, and it avoids the low profitability effect. An advantage over the multiple price procedure is that You cannot influence my choice of \mathcal{D}_2 through Your choice of \mathcal{D}_1 , so \mathcal{D}_1 should not be distorted by an attempt to do so. A disadvantage is that it is no longer clearly in Your interest for \mathcal{D}_1 to avoid sure loss. When \mathcal{D}_1 incurs sure loss there is some choice \mathcal{D}_2 that can exploit it, but this depends on the unknown \mathcal{D}_1 . (See Appendix I4 for examples. This defect could be removed by redefining Your gain Z to be negative whenever Your assessments incur sure loss.) Effects of anticipated information and low demand are still possible, but may be small if my choices provide no extra information about Ω .

The symmetric procedure seems especially suitable for comparing the performance of several probability assessors. An experiment of this kind is described in Appendix I.

How should You choose the assessments x and y in the multiple-price and symmetric procedures? You wish to maximize Your reward Z , which depends on three unknowns u, v, ω as well as x and y . Even if You know u and v , if Your beliefs about ω are indeterminate then there will be many reasonable choices (x, y) , including the *honest* assessments $x = \underline{P}(X)$ and $y = \bar{P}(X)$, and all the precise assessments $x = y$ with $\underline{P}(X) \leq x \leq \bar{P}(X)$. In this case honesty is not uniquely optimal, but it can be defended as a type of minimax action: the honest assessments $x = \underline{P}(X)$ and $y = \bar{P}(X)$ maximize the lower prevision of Your reward, conditional on every pair (u, v) such that $u \leq v$. Thus honesty is a \underline{P} -minimax action (3.9.7) whatever choices are made by Your opponent (provided these avoid sure loss), and it is the only action with this property.

H7 Stakes functions

Let us look more closely at the gambles involved in the multiple-price and symmetric procedures. When the two intervals $[x, y]$ and $[u, v]$ intersect, no gamble is exchanged. Consider the case $v < x$. (The other case $y < u$ is similar.) Define $\lambda(v, x) = \int_v^x \zeta(z) dz$ and $\mu(v, x) = \int_v^x z\zeta(z) dz / \lambda(v, x)$. Then the gamble Z You receive in the multiple-price and symmetric procedures can be written as $Z(v, x) = \int_v^x (X - z)\zeta(z) dz = \lambda(v, x)(X - \mu(v, x))$.

Thus You are effectively buying X for price $\mu(v, x)$, with stake $\lambda(v, x)$. The stake $\lambda(v, x)$ is the integral of the stakes function ζ over the interval (v, x) of mutually acceptable prices. The price $\mu(v, x)$ is the mean of the probability distribution on the interval (v, x) which has density function proportional to ζ . It follows that $v < \mu(v, x) < x$.

For example, the uniform stakes function $\zeta(z) = 1$ has stake $\lambda(v, x) = x - v$, the difference between my selling price and Your buying price, and price $\mu(v, x) = \frac{1}{2}(v + x)$, the arithmetic mean of the two prices.

Your gain Z can also be expressed in terms of the gambles $S(x) = \int_a^x (X - z)\zeta(z) dz$ (where a is a suitable constant), which are called scores. When $v < x$, You receive the gamble $Z(v, x) = S(x) - S(v)$, a difference of scores. In general, You receive

$$Z(u, v, x, y) = \chi(v, x)(S(x) - S(v)) + \chi(y, u)(S(y) - S(u)).$$

H8 Proper scoring rules

Now consider the special case of the multiple-price and symmetric procedures where both You and I specify precise previsions for X , so $x = y$ and

$u = v$. Then Your gain Z from elicitation is

$$Z(v, x) = \chi(v, x)(S(x) - S(v)) + \chi(x, v)(S(x) - S(v)) = S(x) - S(v),$$

which is just the difference between our scores $S(x)$ and $S(v)$. Arranging all the mutually acceptable gambles between two Bayesians is therefore equivalent to comparing their scores, where the scoring rule is defined through the stakes function ζ .

This decomposition of Your gain into the difference of two scores has the important consequence that the various effects of ‘anticipation’ are eliminated. Provided v is chosen independently of x , as in the symmetric procedure, You can influence Your gain Z only through the score $S(x)$. You should therefore choose x to maximize Your expected score. It may be feasible to eliminate the elicitor’s choices, and award You a constant amount plus Your score $S(x)$, so that Your reward depends only on Your assessment x and the true state ω . (The constant amount must be chosen so that the procedure is acceptable to You but not too expensive to the elicitor.)

When Your prevision for X is $P(X)$, the prevision of Your score is $P(S(x)) = \int_a^x (P(X) - z)\zeta(z) dz$. Since ζ is positive, this is maximized by the unique choice $x = P(X)$. It is therefore uniquely optimal to honestly specify Your precise prevision $P(X)$. Scoring rules S with this property are called proper.

The results of Savage (1971) imply that, under mild regularity conditions such as differentiability, scoring rules which are proper for all distributions of X must (essentially) have the form $S(x) = \int_a^x (X - z)\zeta(z) dz$ for some positive stakes function ζ . This means that elicitation using proper scoring rules is (essentially) a special case of the symmetric procedure. For example, the quadratic scoring rule (5.7.6) is defined through the uniform stakes function $\zeta(z) = 1$, while the logarithmic scoring rule (5.12.3(e)) corresponds to $\zeta(z) = z^{-1}$.

Elicitation through scoring rules is certainly simpler than the procedures discussed earlier, because the complicating effect of the elicitor is removed. However, the other difficulties outlined in section H2 still remain. Scoring rules are useful only if Your true previsions are precise (although the symmetric procedure is a useful generalization). Your assessments may be distorted by Your personal interest in ω . Also, You have little incentive to make careful assessments, because Your expected score $P(S(x))$ is insensitive to deviations of x from Your true prevision.

APPENDIX I

The World Cup football experiment

In this appendix we summarize the results of an experiment in which many people assessed upper and lower probabilities for the same events. Similar experiments, in which precise probabilities were elicited, are described by Winkler (1971) and de Fineeti (1972, Chs. 1 and 3).

I1 Description of the experiment

The experiment was designed to elicit upper and lower probabilities concerning all games in the first phase of the 1982 football World Cup, played in Spain, from each of 17 participants. Each participant assessed upper and lower probabilities for the three possible outcomes, ‘win’ (W), ‘draw’ (D), and ‘loss’ (L), in each of 36 games, providing $36 \times 3 \times 2 = 216$ numbers. Most people spent about one or two hours on assessment. (Those who spent longer tended to do better.)

It was explained to the participants that their upper and lower probabilities would be interpreted as minimum selling prices and maximum buying prices for the gambles W, D, L , and that they would be scored by arranging all the mutually acceptable gambles between pairs of participants and accumulating the results. In other words, by specifying upper and lower probabilities, the participants were declaring their willingness to bet at particular rates. Apart from that, no explanation was given of upper and lower probabilities or coherence; one of the aims of the experiment was to see whether naïve assessors would avoid sure loss and be coherent.

The scoring procedure used was the symmetric procedure described in Appendix H6, with uniform stakes function $\zeta(z) = 1$. (The overall ranking of assessors was not sensitive to the choice of stakes function.) Scores were computed as follows. Consider a single outcome W for a particular game, and two assessors who gave (lower, upper) probabilities (x, y) and (u, v) for W . When $v < x$, the two assessors bet on or against W at rate $\frac{1}{2}(x + v)$, with stake $x - v$. They made a similar bet when $y < u$. In the other cases, their probability intervals intersected and no bet on W took place. Formally, let $\chi(v, x)$ take the value one if $v < x$ and zero otherwise. The uncertain gain of

the first assessor from these bets can be written as

$$Z(u, v, x, y) = \chi(v, x)(x - v)(W - \frac{1}{2}(x + v)) + \chi(y, u)(u - y)(\frac{1}{2}(u + y) - W).$$

The overall score for the first assessor was computed by adding together all such gambles Z over all events (W, D, L), all 36 games, and the other 16 probability assessors. At the end of the competition each participant was paid a small amount of money which was a positive linear transformation of his overall score.

Of the 17 participants, 11 were numerate undergraduate students who had no previous training in assessing any kind of subjective probabilities. Three further participants were graduate students with some understanding of the Bayesian approach, and the remaining three were lecturers with a sophisticated knowledge of the Bayesian theory and considerable experience in assessing precise probabilities. Only one of the participants had any theoretical knowledge or practical experience with upper and lower probabilities. The lecturers and graduate students generally scored lower than the undergraduates, apparently because of over-confidence (assigning high probabilities to single outcomes). It seems that a Bayesian education has its disadvantages!

I2 The extent of imprecision

An assessor could refrain from gambling on any game by quoting vacuous upper and lower probabilities, but none of the participants made use of this option. Two of the participants admitted to being almost entirely ignorant about football, but they chose to represent their ignorance through probabilities symmetric in the outcomes W, D, L (e.g. lower probability 0.3 and upper probability 0.4 for each outcome) rather than through high degrees of imprecision $\Delta(A) = \bar{P}(A) - \underline{P}(A)$. These two had overall scores near the average.

At the other extreme, equating upper with lower probabilities would tend to involve the assessor in many bets and increase the amount staked on each game. One participant, a dogmatic Bayesian lecturer, did assign precise probabilities to all events and consequently had much higher total stakes than the others. He finished a distant last overall, although that seems to have been due more to bad judgement about football than to the precision of his assessments; see his eccentric assessments in example I3.

The imprecision in probability assessments was generally greater than had been expected. The value of $\Delta(A)$, averaged over all the assessments made by one person, ranged from 0 (for the Bayesian) to 0.47, with median 0.20. The assessor with highest average imprecision took part in only 294 bets overall (compared with about 700 on average, 1134 for the Bayesian,

and a maximum possible 1728), and the total stakes risked by him were less than one-eighth those of the Bayesian. Thus there was considerable variation in precision amongst participants. There was no clear correlation between an assessor's average degree of imprecision and his overall score.

The precision of probabilities varied across games for most assessors, and most of them stated in an accompanying questionnaire that the precision depended on the amount of information they had about the two teams involved in a game. (Others related the precision to their 'caution', 'sureness' or 'uncertainty' about a game.)

I3 Example of assessments

The assessments of a representative group of nine assessors for the game between Brazil and the USSR are shown in the next table. Here W denotes a win to Brazil. The assessors are numbered according to their overall placings in the competition: assessor 1 finished first and assessor 17 (the Bayesian) finished last. Also tabulated are the total staked by each assessor on the game, which is proportional to the average distance of his probability intervals from those of other assessors, and his accumulated profit on the game resulting from the actual outcome W . (Brazil won the game 2 goals to 1). Those who assigned higher probabilities to W generally (but not necessarily) made higher profits on the game.

Assessor number	$P(W)$	$\bar{P}(W)$	$P(D)$	$\bar{P}(D)$	$P(L)$	$\bar{P}(L)$	Total staked on game	Total profit on game
1	0.67	0.69	0.17	0.20	0.12	0.15	7.7	+2.9
2	0.6	0.9	0.1	0.4	0	0.3	3.2	+1.8
6	0.74	0.79	0.08	0.12	0.14	0.16	10.8	+3.7
7	0.2	0.4	0.3	0.5	0.2	0.4	4.9	-1.0
12	0.4	0.6	0.15	0.25	0.25	0.4	4.2	+1.0
13	0.27	0.52	0.27	0.61	0.21	0.31	3.2	-0.2
14	0.5	0.9	0.35	0.6	0.1	0.6	3.0	+0.8
16	0.60	0.65	0.20	0.25	0.35	0.40	7.3	+1.4
17	0.1	0.1	0.8	0.8	0.1	0.1	25.1	-13.7

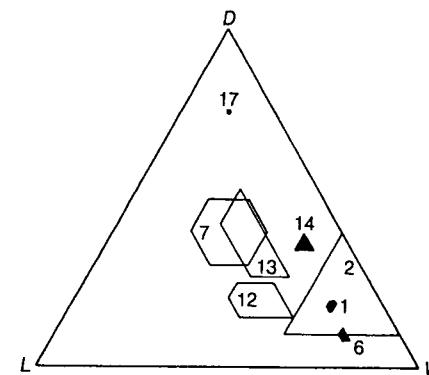


Figure I3 Probability models of eight assessors for the game Brazil vs USSR.

These assessments can be compared most easily by plotting $\mathcal{M}(P)$ for each assessor as a convex polygon in the probability triangle (section 4.2.3). The polygon for assessor 13 was constructed in section 4.6.1 from his six probability assessments, which determine six lines parallel to the sides of the triangle, and the other polygons can be constructed in the same way. They are shown in Figure I3, each labelled with its assessor number.

I4 Avoiding sure loss

The class \mathcal{M} is empty just when the six upper and lower probabilities incur sure loss. All the assessments satisfied $0 \leq P(A) \leq \bar{P}(A) \leq 1$ for each event A . In that case, avoiding sure loss is equivalent to the condition $P(W) + P(D) + P(L) \leq 1 \leq \bar{P}(W) + \bar{P}(D) + \bar{P}(L)$.

Assessor 16 incurred sure loss in the tabulated game, since the sum of his lower probabilities (0.60, 0.20, 0.35) is greater than one. If that had been known to other assessors, they could have chosen their own probabilities to exploit it. By specifying precise probabilities $P(W) = 0.55$, $P(D) = 0.15$, $P(L) = 0.30$, You could stake 0.05 with assessor 16 against each of the outcomes W, D, L at rates 0.575, 0.175, 0.325 respectively, from which You are sure to gain $0.05 \times 0.075 = 0.00375$. Of course the 'sure loss' could not be exploited since assessments were not revealed to other assessors. Anyway, the amount is minute in comparison to actual profits on the game.

Despite the 'sure loss', assessor 16 actually made a profit on this game, largely due to his high lower probability for W . Indeed, on another game the same assessor incurred sure loss, according to our definition, but nevertheless achieved a 'sure gain' in the sense that he would have made a profit on the game whichever of W, D, L occurred! This shows that, in practice, 'incurring sure loss' is not necessarily harmful.

Assessor 16 violated the constraint $\underline{P}(W) + \underline{P}(D) + \underline{P}(L) \leq 1$ for every game. He finished next to last in the competition, although still a long way ahead of the Bayesian. It is difficult to tell to what extent his poor performance was due to incurring sure loss.

The constraint $\underline{P}(W) + \underline{P}(D) + \underline{P}(L) \leq 1$ was violated for only one game each by four other participants, and it seems likely that these were unintended slips. In no case was the constraint $\bar{P}(W) + \bar{P}(D) + \bar{P}(L) \geq 1$ violated. This is evidence that naïve assessors of upper and lower probabilities do tend to avoid sure loss, at least in simple problems.

I5 Coherence

The six probability assessments are coherent just when each of the six corresponding lines touches the polygon \mathcal{M} . Coherence is equivalent to the condition $\underline{P}(A) + \underline{P}(B) + \bar{P}(C) \leq 1 \leq \bar{P}(A) + \bar{P}(B) + \bar{P}(C)$ whenever (A, B, C) is a permutation of (W, D, L) . Of the assessors listed in the table, only numbers 13, 14 and 16 are incoherent. Assessor 16 is incoherent because he incurred sure loss, assessor 14 violated each of the three constraints $\underline{P}(A) + \underline{P}(B) + \bar{P}(C) \leq 1$, and assessor 13 violated one of these constraints (see section 4.6.1).

Coherence was violated quite often in the experiment. Only three assessors managed to be completely coherent, and one of these was the Bayesian. The others ranged from a few violations to violations in every game. About 90% of the violations were of constraints $\underline{P}(A) + \underline{P}(B) + \bar{P}(C) \leq 1$, and many of these had $\underline{P}(A) + \underline{P}(B) + \bar{P}(C) > 1.2$, whereas there were few cases of $\underline{P}(A) + \bar{P}(B) + \bar{P}(C) < 0.9$.

The effect of incoherence was to make assessors more cautious in gambling than they would have been if they had recognized the incoherence and removed it by natural extension. Assessor 14, for example, has high degree of imprecision and low total stake, but his polygon \mathcal{M} is much smaller than these suggest. Those who were most often incoherent tended to have low total stakes, but on the whole their scores were no worse than those who were coherent. (Assessor 16, who incurred sure loss, is an exception.)

I6 Natural extension

In terms of the simplex representation, natural extension of P corresponds to replacing P by the envelope of the polygon $\mathcal{M}(P)$. The natural extensions to coherent upper and lower probabilities are simply given by $\bar{E}(A) = \min\{\bar{P}(A), 1 - \underline{P}(B) - \bar{P}(C)\}$ and $\underline{E}(A) = \max\{\underline{P}(A), 1 - \bar{P}(B) - \bar{P}(C)\}$. For instance, natural extension greatly modifies the upper probabilities for assessor 14, to $\bar{E}(W) = 0.55$, $\bar{E}(D) = 0.4$, $\bar{E}(L) = 0.15$.

The polygon $\mathcal{M}(P)$ can have any number of extreme points between one

and six. Figure I3 has examples of all except two extreme points, which occurs when precise probabilities $\underline{P}(A) = \bar{P}(A)$ are assessed for just one event A . The length of each side of the polygon is proportional to the amount of slack in one of the coherence constraints, e.g. the length of the side corresponding to $\bar{P}(W)$ is proportional to $1 - \bar{P}(W) - \underline{P}(D) - \bar{P}(L)$. One of the six possible sides of the polygon is absent just when there is zero or negative slack in the corresponding constraint. (Negative slack indicates incoherence.) For example, the polygon for assessor 13 has four sides because he violates one coherence constraint and has zero slack in another.

I7 How \mathcal{M} affects scores

Bets were arranged between two assessors i and j whenever their probability intervals were disjoint. Assuming that each assessor was coherent, it is easy to see that they bet with each other on a game just when their polygons \mathcal{M}_i and \mathcal{M}_j were disjoint. Geometrically, if \mathcal{M}_i and \mathcal{M}_j are disjoint then there is a line parallel to one of the sides of the probability triangle that separates them.

From Figure I3 we see that, on the game Brazil vs USSR, assessor 2 bet with assessors 7 (on W) and 17 (on W and against D) but not with assessors 1, 6 and 12. (All these were coherent.)

The total amount staked by an assessor on a game is proportional to the average distance of his probability intervals from the intervals of other assessors. This depends on both the precision of the intervals and their position relative to other intervals. For coherent assessors the precision of assessments can be measured by the area of the polygon \mathcal{M} . Assessors with larger polygons tend to bet with fewer people and have lower total stake. For example, the assessments of 1 have much greater precision than those of 2, and that is reflected in 1's larger total stake.

The stake also depends on the position of \mathcal{M} relative to the other polygons. Outlying assessments produce higher stakes, e.g. assessors 1 and 6 have similar precision but 6 has higher total stake because his assessments concerning W and D disagree more with the other assessors. The assessments of assessor 17 are both outlying and maximally precise, producing a very high total stake and exceptionally high loss on the game.

It was possible, by choosing \mathcal{M} in the centre of the other assessments, to obtain a **sure gain**, that is, to be guaranteed a net profit on a game, irrespective of the outcome W, D or L . That happens when the assessor manages to balance bets on the outcome A with bets against A , for each A . Bookmakers routinely adjust their upper and lower probabilities in the light of the demand for bets to achieve this, as suggested in Appendix H4, but in this experiment it happened without knowledge of the demand for bets. Sure gains were

quite frequent, achieved by an average of nearly one person per game, although there were none in the tabulated game.

There was a strong correlation in the experiment between overall scores and the average degree of uniformity of probabilities, as measured (for example) by their upper entropy (note 14 of section 5.12). Those whose polygons \mathcal{M} were generally near the centre of the probability triangle, indicating similar probabilities for the three outcomes W, D and L , tended to do better. This kind of correlation was observed by Winkler (1971).

APPENDIX J

Regular extension

We saw in section 6.8 that whenever $\underline{P}(B) = 0$, the minimal conditional previsions $\underline{P}(\cdot|B)$ that are coherent with \underline{P} are vacuous. Apparently, \underline{P} provides no information about updated previsions after observing B . When $\bar{P}(B) > 0$, however, some information can be recovered from \underline{P} and used to define non-vacuous conditional previsions $\underline{R}(\cdot|B)$, called the regular extension.

J1 Definition

Suppose \underline{P} is a coherent lower prevision on $\mathcal{L}(\Omega)$, and B is a subset of Ω . Define the **regular extension** $\underline{R}(\cdot|B)$ to be the vacuous conditional prevision when $\bar{P}(B) = 0$, and

$$\underline{R}(X|B) = \max \{\mu: \underline{P}(B(X - \mu)) \geq 0\} \quad \text{when } \bar{P}(B) > 0.$$

When $\bar{P}(B) > 0$, $\underline{R}(\cdot|B)$ is just the lower envelope of the class of linear conditional previsions obtained from $\mathcal{M}(\underline{P})$ by eliminating P such that $P(B) = 0$, and applying Bayes' rule to the others. That is,

$$\underline{R}(X|B) = \inf \{P(BX)/P(B): P \in \mathcal{M}(\underline{P}), P(B) > 0\}.$$

(To see that, note that $\underline{P}(B(X - \mu)) \geq 0$ just when $P(BX) - \mu P(B) \geq 0$ for all $P \in \mathcal{M}(\underline{P})$. Those P with $P(B) = 0$ are irrelevant because they always satisfy this condition.)

J2 Comparison with natural extension

By Theorem 6.8.2, the natural extension $\underline{E}(X|B) = \mu$ is defined by the GBR $\underline{P}(B(X - \mu)) = 0$ when $\underline{P}(B) > 0$, and is otherwise vacuous. The regular extension therefore agrees with the natural extension unless $\underline{P}(B) = 0$ and $\bar{P}(B) > 0$. In that case, the regular extension is typically non-vacuous and can be relatively precise. For example, if only one extreme point of $\mathcal{M}(\underline{P})$ has $P(B) > 0$, then $\underline{R}(\cdot|B)$ is a linear prevision but $\underline{E}(\cdot|B)$ is vacuous. The difference between regular and natural extension is also illustrated by examples J4 and J6.

J3 Coherence

Suppose now that $\underline{R}(\cdot|B)$ is defined for all sets B in a partition \mathcal{B} . Because $\underline{R}(\cdot|\mathcal{B})$ are lower envelopes of linear previsions, they are separately coherent. The regular extensions $\underline{R}(\cdot|\mathcal{B})$ are coherent with \underline{P} if and only if \underline{P} satisfies the regularity condition

(P9) if $X \in \mathcal{L}$, and (for all $B \in \mathcal{B}$) $\underline{P}(BX) \geq 0$ when $\bar{P}(B) > 0$ and $BX = 0$ otherwise, then $\underline{P}(X) \geq 0$.

(To prove this, verify that \underline{P} and $\underline{R}(\cdot|\mathcal{B})$ must satisfy C12 of section 6.5.3, and show that C11 is equivalent to P9.)

This regularity axiom is similar to, but stronger than, the \mathcal{B} -conglomerability axiom P7. It can be defended as a rationality condition provided we regard a gamble Y as desirable whenever $\underline{P}(Y) \geq 0$ and $\bar{P}(Y) > 0$. (Then every gamble $B(X + \delta)$ is desirable in P9, so $X + \delta$ should be desirable, giving $\underline{P}(X) \geq 0$.)

This suggests strengthening the axioms for coherence of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ by requiring also

(C16) if $X \in \mathcal{L}, B \in \mathcal{B}, \bar{P}(B) > 0$ and $\underline{P}(BX) \geq 0$ then $\underline{P}(X|B) \geq 0$.

This axiom is clearly satisfied by \underline{P} and $\underline{R}(\cdot|\mathcal{B})$. In fact, there is some conditional prevision $\underline{P}(\cdot|\mathcal{B})$ that is coherent with \underline{P} and also satisfies C16 if and only if \underline{P} satisfies P9. Whenever $\bar{P}(B) > 0$, the regular extension $\underline{R}(\cdot|B)$ is the unique coherent extension that satisfies C16.

J4 Example

Any separately coherent conditional prevision $\underline{P}(\cdot|\mathcal{B})$ can be obtained as the regular extension of some lower prevision \underline{P} . That is, any $\underline{P}(\cdot|\mathcal{B})$ can be obtained by ‘conditioning’ some \underline{P} . Simply take \underline{P} to be the natural extension of $\underline{P}(\cdot|\mathcal{B})$, defined in Theorem 6.7.2 as $\underline{P}(X) = \inf \underline{P}(X|\mathcal{B})$. Then \underline{P} has vacuous \mathcal{B} -marginal, so $\underline{P}(B) = 0$ and $\bar{P}(B) = 1$ for every B in \mathcal{B} . Since $\underline{P}(B(X - \mu)) = \min\{0, \underline{P}(X|B) - \mu\}$, the regular extension of \underline{P} is $\underline{R}(X|B) = \max\{\mu : \underline{P}(B(X - \mu)) \geq 0\} = \max\{\mu : \underline{P}(X|B) - \mu \geq 0\} = \underline{P}(X|B)$. Thus the regular extension of \underline{P} is $\underline{P}(\cdot|\mathcal{B})$. Here \underline{P} satisfies P9, so \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent. In fact, $\underline{P}(\cdot|\mathcal{B})$ is the unique conditional prevision that is coherent with \underline{P} and satisfies C16.

J5 Regular posteriors

Consider a statistical problem with parameter θ , prior prevision \underline{P} , observation x and precise likelihood function L_x . As in section 8.4.4, the unique coherent extension \underline{E} of the prior and likelihood satisfies

$\underline{E}(\{x\}(Y - \mu)) = \underline{P}(L_x(Y - \mu))$ when Y is a function of θ . Hence the regular extension of \underline{E} is $\underline{R}(Y|x) = \max\{\mu : \underline{P}(L_x(Y - \mu)) \geq 0\}$, provided $\bar{P}(L_x) > 0$. This $\underline{R}(\cdot|x)$ is called the **regular posterior**. It is the lower envelope of Bayesian posteriors for linear priors in $\mathcal{M}(\underline{P})$ that assign positive probability to x , provided there is such a prior. It depends on the sampling model only through the observed likelihood function L_x .

J6 Beta–Bernoulli example

Consider the near-ignorance prior \underline{P} defined in section 5.3.2, which is the lower envelope of beta (s_0, t) distributions over $0 < t < 1$. When L_x is the Bernoulli likelihood, $\underline{P}(L_x) = 0$ for all possible x , so the posteriors defined by natural extension are vacuous. (Compare with Example 8.4.9.) However, $\bar{P}(L_x) > 0$ for all beta distributions P , so the regular posteriors $\underline{R}(\cdot|x)$ are the lower envelopes of beta posteriors, as defined in section 5.3.1. These are the unique posteriors that are coherent with the near-ignorance prior and Bernoulli sampling model and also satisfy axiom C16. The regular posteriors are nearly precise when the sample is large.

APPENDIX K

W-coherence

When can a coherent collection of lower previsions $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ be written as lower envelopes of a class Γ of coherent linear collections $P_\gamma(\cdot|\mathcal{B}_1), \dots, P_\gamma(\cdot|\mathcal{B}_m)$? We know from Theorem 7.1.6 that the lower envelopes of any class of coherent linear collections are coherent. We know also, from various examples in sections 6.6 and 6.9, that the converse fails: there are coherent collections of lower previsions that are not dominated by any coherent collection of linear previsions.

Compare these examples with the earlier result (Theorem 3.3.3) that every coherent lower prevision \underline{P} is the lower envelope of some class of linear previsions. This result, and the sensitivity analysis interpretation it supports, do not extend to the general case of conditional previsions.

Theorem 3.3.3 does generalize, however, to the case in which all the partitions \mathcal{B}_i are finite. That follows from the results of Williams (1975a), who defines a concept of coherence that we will call W-coherence. (Regazzini, 1983, adopts an equivalent definition in the case of linear previsions.) In general, W-coherence is weaker than our concept of coherence (7.1.4(b)), but it is equivalent to coherence when all the partitions are finite.

K1 Definition

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ are defined on linear spaces and separately coherent, for $1 \leq i \leq m$. Call them **W-coherent** if they satisfy the coherence condition 7.1.4(b) whenever each $S_i(Y_i)$ is a finite set.

In other words, condition 7.1.4(b) is required to hold just when it involves only finitely many contingent gambles $G(Y_i|B)$. Coherence clearly implies W-coherence. Because the support $S_i(Y_i)$ can be any subset of \mathcal{B}_i , coherence is stronger than W-coherence when some \mathcal{B}_i is infinite, but the two conditions are equivalent when each \mathcal{B}_i is finite.

K2 Single partition

Suppose, for example, that an unconditional prevision \underline{P} and a single conditional prevision $\underline{P}(\cdot|\mathcal{B})$ are specified. Theorem 6.5.4 shows that, for

finite \mathcal{B} and suitable restrictions on the domains, coherence of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ reduces to the generalized Bayes rule C12. In fact, W-coherence is equivalent to C12 more generally, for infinite partitions \mathcal{B} as well as finite ones, whereas the extra conglomerative axiom C11 is needed for coherence. When the unconditional prevision is linear, W-coherence is equivalent to Bayes' rule. It follows that the linear previsions in Examples 6.6.6 and 6.6.7 are W-coherent, although they incur sure loss.

When $\bar{P}(B)=0$ for all B in \mathcal{B} , as will typically hold for uncountable partitions \mathcal{B} , C12 holds trivially. Then W-coherence of \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ requires no more than coherence of \underline{P} and separate coherence of $\underline{P}(\cdot|\mathcal{B})$. These results indicate that W-coherence is considerably weaker than coherence when some partition is infinite.

Williams (1975a, Theorem 2) proved that all W-coherent collections of lower previsions are lower envelopes of W-coherent linear collections. Hence we obtain the following converse to Theorem 7.1.6.

K3 Williams' theorem

Suppose that $\underline{P}(\cdot|\mathcal{B}_i)$ is defined on a linear space \mathcal{K}_i and \mathcal{B}_i is finite, for $1 \leq i \leq m$. Then $\underline{P}(\cdot|\mathcal{B}_1), \dots, \underline{P}(\cdot|\mathcal{B}_m)$ are coherent if and only if they are lower envelopes of a class of coherent linear collections, i.e. if and only if there are coherent linear previsions $P_\gamma(\cdot|\mathcal{B}_1), \dots, P_\gamma(\cdot|\mathcal{B}_m)$ for each $\gamma \in \Gamma$ such that $\underline{P}(Y|B) = \inf\{P_\gamma(Y|B): \gamma \in \Gamma\}$ for all $B \in \mathcal{B}_i$, $Y \in \mathcal{K}_i$ and $1 \leq i \leq m$.

For example, suppose that \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ are coherent, where \mathcal{B} is finite. Then Γ can be taken to be the class of all coherent linear pairs P , $P(\cdot|\mathcal{B})$ such that P dominates \underline{P} and $P(\cdot|\mathcal{B})$ dominates $\underline{P}(\cdot|\mathcal{B})$. When the domain of \underline{P} is sufficiently large and $\underline{P}(B) > 0$ for all B in \mathcal{B} , $\underline{P}(\cdot|\mathcal{B})$ is determined by \underline{P} through the generalized Bayes rule and the index set Γ can be identified with $\mathcal{M}(\underline{P})$. We then have $\underline{P}(Y) = \min\{P(Y): P \in \mathcal{M}(\underline{P})\}$, $\underline{P}(Y|B) = \min\{P(Y|B)/P(B): P \in \mathcal{M}(\underline{P})\}$.

K4 Infinite partitions

The applicability of Williams' theorem is limited by the assumption that all \mathcal{B}_i are finite. In statistical problems, that holds only when the parameter space and sample space are both finite, which is atypical. In the case of infinite \mathcal{B}_i , the theorem implies that coherent collections are always lower envelopes of W-coherent linear collections, although the latter need not be coherent. Consider Example 6.6.9. There the lower prevision \underline{P} was constructed as the lower envelope of linear previsions P_1 and P_2 . These determine, through Bayes' rule, linear conditional previsions $P_1(\cdot|\mathcal{B})$ and $P_2(\cdot|\mathcal{B})$ whose lower envelope is the vacuous conditional prevision $\underline{P}(\cdot|\mathcal{B})$.

Thus the coherent pair $\underline{P}, \underline{P}(\cdot|\mathcal{B})$ is the lower envelope of the linear pairs $P_1, P_1(\cdot|\mathcal{B})$ and $P_2, P_2(\cdot|\mathcal{B})$. Both linear pairs are W-coherent since they satisfy Bayes' rule, but we saw in Example 6.6.9 that neither linear pair is coherent. In this case \underline{P} and $\underline{P}(\cdot|\mathcal{B})$ have no dominating linear pairs that are coherent.

W-coherence is a natural generalization of unconditional coherence, especially under a sensitivity analysis interpretation. It has the advantage that W-coherent models can always be extended W-coherently to larger domains (cf. section 8.1). Because it involves only finitely many contingent gambles and does not rely on the conglomerative principle, W-coherence is also a natural generalization of de Finetti's (1974) definition of coherence. Nevertheless, W-coherence seems too weak a requirement when the partitions are infinite, and the stronger conditions 6.3.2 and 7.1.4 seem compelling.

References

-
- Aczel, J. (1966). *Lectures on Functional Equations and their Applications*. Academic Pr., New York.
- Aczel, J. and Daroczy, Z. (1975). *On Measures of Information and their Characterizations*. Academic Pr., New York.
- Adams, J.B. (1976). A probability model of medical reasoning and the MYCIN model. *Math. Biosciences* **32**, 177–186. Reprinted in Buchanan and Shortliffe (1984).
- Agassi, J. (1975). Subjectivism: From infantile disease to chronic illness (with discussion). *Synthese* **30**, 3–38.
- Allais, M. and Hagen, O. (eds.) (1979). *Expected Utility Hypotheses and the Allais Paradox*. Reidel, Dordrecht.
- Amaral-Turkman, M.A. and Dunsmore, I.R. (1985). Measures of information in the predictive distribution. In *Bayesian Statistics 2*, 603–612, ed. J.M. Bernardo *et al.*
- Anderson, J.R. (1980). *Cognitive Psychology and Its Implications*. Freeman, San Francisco.
- Andrews, S.E. (1989). *Practical computational methods for aggregation of lower envelopes*. Ph.D. Dissertation, Cornell U., Ithaca (NY).
- Anger, B. (1977). Representation of capacities. *Math. Annalen* **229**, 245–258.
- Anscombe, F.J. and Aumann, R.J. (1963). A definition of subjective probability. *Ann. Math. Statist.* **34**, 199–205.
- Aristotle (1962). *Nicomachean Ethics*. Translation by M. Ostwald. Bobbs-Merrill, Indianapolis.
- Armstrong, D.M. (1973). *Belief, Truth and Knowledge*. Cambridge U. Pr.
- Armstrong, T.E. (1987). A coherent view of comparative probability. *Research Report 87–05*, Math. Dept., U. Maryland, Baltimore County.
- Aumann, R.J. (1962). Utility theory without the completeness axiom. *Econometrica* **30**, 445–462, and **32**, 210–212.
- Ayer, A.J. (ed.) (1959). *Logical Positivism*. Free Pr., New York.
- Ayer, A.J. (1972). *Probability and Evidence*. Macmillan, London.
- Ayer, A.J. (1973). *The Central Questions of Philosophy*. Penguin, Middlesex.

- Barnard, G.A. (1967). The use of the likelihood function in statistical practice. In *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* 1, 27–40. U. California Pr., Berkeley.
- Barnard, G.A. (1980). Pivotal inference and the Bayesian controversy (with discussion). In *Bayesian Statistics 1*, 295–318, ed. J.M. Bernardo *et al.*
- Barnard, G.A., Jenkins, G.M. and Winsten, C.B. (1962). Likelihood, inference, and time series. *J. Roy. Statist. Soc. Ser. A* 125, 321–372.
- Barnard, G.A. and Sprott, D.A. (1983). The generalised problem of the Nile: robust confidence sets for parametric functions. *Ann. Statist.* 11, 104–113.
- Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika* 67, 293–310.
- Barnett, V. (1982). *Comparative Statistical Inference*, 2nd edition. Wiley, London.
- Baron, J. (1985). *Rationality and Intelligence*. Cambridge U. Pr.
- Basu, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā Ser. A* 37, 1–71.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc. London* A53, 370–418. Reprinted in *Biometrika* 45 (1958), 293–315.
- Beach, B.H. (1975). Expert judgement about uncertainty: Bayesian decision making in realistic settings. *Organiz. Behav. Hum. Perform.* 14, 10–59.
- Becker, S.W. and Brownson, F.O. (1964). What price ambiguity? Or the role of ambiguity in decision making. *J. Political Economy* 72, 62–73.
- Bellman, R.E. and Giertz, M. (1973). On the analytic formalism of the theory of fuzzy sets. *Inform. Sci.* 5, 149–156.
- Bellman, R.E. and Zadeh, L.A. (1970). Decision-making in a fuzzy environment. *Management Sci.* 17B, 141–164.
- Beran, R.J. (1971). On distribution-free statistical inference with upper and lower probabilities. *Ann. Math. Statist.* 42, 157–168.
- Berberian, S.K. (1974). *Lectures in Functional Analysis and Operator Theory*. Springer-Verlag, New York.
- Berger, J.O. (1980). *Statistical Decision Theory: Foundations, Concepts, and Methods*. Springer-Verlag, New York.
- Berger, J.O. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses*, ed. J.B. Kadane. Elsevier Science, Amsterdam.
- Berger, J.O. (1985a). *Statistical Decision Theory and Bayesian Analysis*. (Second edition of Berger (1980).) Springer-Verlag, New York.
- Berger, J.O. (1985b). In defense of the likelihood principle: axiomatics and coherency (with discussion). In *Bayesian Statistics 2*, 33–65, ed. J.M. Bernardo *et al.*
- Berger, J.O. (1985c). The frequentist viewpoint and conditioning. In *Procs.*

- Berkeley Conf. in Honor of Kiefer and Neyman*, eds. L. LeCam and R. Olshen. Wadsworth, Belmont (CA).
- Berger, J.O. and Berliner, L.M. (1986). Robust Bayes and empirical Bayes analysis with ϵ -contaminated priors. *Ann. Statist.* 14, 461–486.
- Berger, J.O. and Wolpert, R.L. (1984). *The Likelihood Principle*. Volume 6 of IMS Lecture Notes – Monograph Series. Inst. Math. Statist., Hayward (CA).
- Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* 41, 113–147.
- Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M. (eds.) (1980). *Bayesian Statistics 1*. Valencia U. Pr.
- Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M. (eds.) (1985). *Bayesian Statistics 2*. North-Holland, Amsterdam.
- Bernoulli, J. (1713). *Ars Conjectandi, Pars Quarta*. Basel. English translation by Bing Sung, Tech. Report 12, Dept. of Statistics, Harvard U.
- Beyth-Marom, R. (1982). How probable is probable? Numerical translations of verbal probability expressions. *J. Forecasting* 1, 257–269.
- Bhaskara Rao, K.P.S. and Bhaskara Rao, M. (1983). *Theory of Charges*. Academic Pr., London.
- Billingsley, P. (1979). *Probability and Measure*. Wiley, New York.
- Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* 57, 269–326.
- Birnbaum, A. (1969). Concepts of statistical evidence. In *Philosophy, Science and Method*, eds. S. Morgenbesser, P. Suppes and M. White. St. Martin's Pr., New York.
- Birnbaum, A. (1977). The Neyman–Pearson theory as decision theory and as inference theory: with a criticism of the Lindley–Savage argument for Bayesian theory. *Synthese* 36, 19–49.
- Blachman, N.M. (1966). *Noise and its Effect on Communication*. McGraw-Hill, New York.
- Black, M. (1970). *Margins of Precision*. Cornell U. Pr., Ithaca (NY).
- Blackburn, S. (1980). Opinions and chances. In *Prospects for Pragmatism*, ed. D.H. Mellor. Cambridge U. Pr.
- Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* 33, 882–886.
- Block, N. (ed.) (1980). *Readings in Philosophy of Psychology*, Vol. 1. Harvard U. Pr.
- Blum, J.R. and Rosenblatt, J. (1967). On partial a priori information in statistical inference. *Ann. Math. Statist.* 38, 1671–1678.
- Boesky, I.F. (1985). *Merger Mania*. Holt, Rinehart and Winston, New York.
- Bondar, J.V. (1977). A conditional confidence principle. *Ann. Statist.* 5, 881–891.

REFERENCES

- Boole, G. (1854). *An Investigation of The Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities*. Macmillan, London.
- Borel, E. (1924). A propos d'un traité de probabilités. *Revue Philosophique* 98, 321–336. English translation in Kyburg and Smokler (1964).
- Borel, E. (1943). *Les Probabilités et la Vie*. Pr. Universitaires de France. English translation *Probabilities and Life* (1962), Dover, New York.
- Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics*, 201–236, eds. R.L. Launer and G.N. Wilkinson. Academic Pr., New York.
- Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* 143, 383–430.
- Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). *Statistics for Experimenters*. Wiley, New York.
- Box, G.E.P. and Tiao, G.C. (1962). A further look at robustness via Bayes' theorem. *Biometrika* 49, 419–432.
- Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading (MA).
- Braithwaite, R.B. (1946). Belief and action. *Proc. Aristot. Soc., Suppl.* Vol. 20, 1–19.
- Braithwaite, R.B. (1953). *Scientific Explanation*. Cambridge U. Pr.
- Breiman, L. (1968). *Probability*. Addison-Wesley, Reading (MA).
- Bridgman, P.W. (1927). *The Logic of Modern Physics*. Macmillan, New York.
- Brown, L.D. (1967). The conditional level of Student's *t* test. *Ann. Math. Statist.* 38, 1068–1071.
- Brown, L.D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* 6, 59–71.
- Buchanan, B.G. and Shortliffe, E.H. (eds.) (1984). *Rule-Based Expert Systems*. Addison-Wesley, Reading (MA).
- Budescu, D.V., Weinberg, S. and Wallsten, T.S. (1988). Decisions based on numerically and verbally expressed uncertainties. *J. Exper. Psychology: Human Perception and Performance* 14, 281–294.
- Buehler, R.J. (1959). Some validity criteria for statistical inferences. *Ann. Math. Statist.* 30, 845–863.
- Buehler, R.J. (1976). Coherent preferences. *Ann. Statist.* 4, 1051–1064.
- Buehler, R.J. (1982). Some ancillary statistics and their properties (with discussion). *J. Amer. Statist. Assoc.* 77, 581–594.
- Buehler, R.J. and Feddersen, A.P. (1963). Note on a conditional property of Student's *t*. *Ann. Math. Statist.* 34, 1098–1100.
- Burgess, J.P. (1969). Probability logic. *J. Symbolic Logic* 34, 264–274.
- Cano, J.A., Hernandez, A. and Moreno, E. (1985). Posterior measures under partial prior information. *Statistica* 2, 219–230.

REFERENCES

- Carnap, R. (1936–37). Testability and meaning. *Philos. Sci.* 3, 419–471, and 4, 1–40.
- Carnap, R. (1952). *The Continuum of Inductive Methods*. U. Chicago Pr.
- Carnap, R. (1956). The methodological character of theoretical concepts. In *Minnesota Studies in the Philosophy of Science*, Vol. I, 38–76, eds. H. Feigl and M. Scriven. Minneapolis.
- Carnap, R. (1962). *Logical Foundations of Probability*, 2nd edition. U. Chicago Pr.
- Carnap, R. (1971). Inductive logic and rational decisions. Chapter 1 of *Studies in Inductive Logic and Probability*, Vol. 1, eds. R. Carnap and R.C. Jeffrey. U. California Pr., Berkeley.
- Carnap, R. (1980). A basic system of inductive logic, part 2. In *Studies in Inductive Logic and Probability*, Vol. 2, 7–155, ed. R.C. Jeffrey. U. California Pr., Berkeley.
- Casella, G. (1986). Conditionally acceptable frequentist solutions. *Report BU-908-M*, Biometrics Unit, Cornell U., Ithaca (NY).
- Chamberlain, G. and Leamer, E.E. (1976). Matrix weighted averages and posterior bounds. *J. Roy. Statist. Soc. Ser. B* 38, 73–84.
- Chernoff, H. and Moses, L.E. (1959). *Elementary Decision Theory*. Wiley, New York.
- Choquet, G. (1953–54). Theory of capacities. *Ann. Inst. Fourier (U. Grenoble)* 5, 131–295.
- Cohen, L.J. (1977). *The Probable and the Provable*. Clarendon Pr., Oxford.
- Cohen, L.J. (1981). Can human irrationality be experimentally demonstrated? (with discussion). *Behav. Brain Sci.* 4, 317–370.
- Cornfield, J. (1969). The Bayesian outlook and its application (with discussion). *Biometrics* 25, 617–657.
- Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* 29, 357–372.
- Cox, D.R. (1971). The choice between alternative ancillary statistics. *J. Roy. Statist. Soc. Ser. B* 33, 251–255.
- Cox, D.R. (1980). Local ancillarity. *Biometrika* 67, 279–286.
- Cox, D.R. (1988). Some aspects of conditional and asymptotic inference: a review. *Sankhyā, Ser. A* 50, 314–337.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
- Cox, D.R. and Hinkley, D.V. (1978). *Problems and Solutions in Theoretical Statistics*. Chapman and Hall, London.
- Cox, D.R. and Snell, E.J. (1981). *Applied Statistics*. Chapman and Hall, London.
- Cox, R.T. (1946). Probability, frequency, and reasonable expectation. *Amer. J. Phys.* 14, 1–13.

- Cox, R.T. (1961). *The Algebra of Probable Inference*. Johns Hopkins Pr., Baltimore.
- Curley, S.P. and Yates, J.F. (1985). The center and range of the probability interval as factors affecting ambiguity preferences. *Organiz. Behav. and Human Decision Processes* **36**, 273–287.
- Dalal, S.R. and Hall, W.J. (1983). Approximating priors by mixtures of natural conjugate priors. *J. Roy. Statist. Soc. Ser. B* **45**, 278–286.
- Daniell, P.J. (1917–18). A general form of integral. *Ann. Math.* **19**, 279–294.
- Darwall, S.L. (1983). *Impartial Reason*. Cornell U. Pr., Ithaca (NY).
- Davis, M. (1977). *Applied Nonstandard Analysis*. Wiley, New York.
- Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41**, 1–31.
- Dawid, A.P. (1982a). Intersubjective statistical models. In *Exchangeability in Probability and Statistics*, eds. G. Koch and F. Spizzichino. North-Holland, Amsterdam.
- Dawid, A.P. (1982b). Probability, symmetry and frequency. *Research Report* 13, Dept. of Statistical Sci., University College, London.
- Dawid, A.P. (1982c). The well-calibrated Bayesian (with discussion). *J. Amer. Statist. Assoc.* **77**, 605–613.
- Dawid, A.P. (1986). A Bayesian view of statistical modelling. In *Bayesian Inference and Decision Techniques*, eds. P.K. Goel and A. Zellner. North-Holland, Amsterdam.
- Dawid, A.P. and Stone, M. (1972). Expectation consistency of inverse probability distributions. *Biometrika* **59**, 486–489.
- Dawid, A.P. and Stone, M. (1973). Expectation consistency and generalized Bayes inference. *Ann. Statist.* **1**, 478–485.
- Dawid, A.P. and Stone, M. (1982). The functional-model basis of fiducial inference (with discussion). *Ann. Statist.* **10**, 1054–1074.
- Dawid, A.P., Stone, M. and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. Ser. B* **35**, 189–233.
- Day, J.P. (1961). *Inductive Probability*. Routledge and Kegan Paul, London.
- Day, M.M. (1942). Ergodic theorems for Abelian semigroups. *Trans. Amer. Math. Soc.* **51**, 399–412.
- Day, M.M. (1973). *Normed Linear Spaces* (3rd edition). Springer-Verlag, Berlin.
- De Morgan, A. (1847). *Formal Logic*. Taylor and Walton, London.
- DeGroot, M.H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.
- Dempster, A.P. (1964). On the difficulties inherent in Fisher's fiducial argument. *J. Amer. Statist. Assoc.* **60**, 420–436.
- Dempster, A.P. (1966). New methods for reasoning towards posterior

- distributions based on sample data. *Ann. Math. Statist.* **37**, 355–374.
- Dempster, A.P. (1967a). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **38**, 325–339.
- Dempster, A.P. (1967b). Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika* **54**, 515–528.
- Dempster, A.P. (1968). A generalization of Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **30**, 205–247.
- Dempster, A.P. (1969). Upper and lower probability inferences for families of hypotheses with monotone density ratios. *Ann. Math. Statist.* **40**, 953–969.
- Dempster, A.P. (1975). A subjectivist look at robustness. *Bull. Internat. Statist. Inst.* **46**, 349–374.
- Dempster, A.P. (1985). Probability, evidence, and judgment (with discussion). In *Bayesian Statistics 2*, 119–131, ed. J.M. Bernardo *et al.*
- Dennett, D.C. (1978). *Brainstorms*. Bradford Books, Montgomery (VT).
- Dennett, D.C. (1984). *Elbow Room*. MIT Pr., Cambridge (MA).
- Dennett, D.C. (1987). *The Intentional Stance*. MIT Pr., Cambridge (MA).
- DeRobertis, L. and Hartigan, J.A. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9**, 235–244.
- Dewey, J. (1933). *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process*. Heath, Boston.
- Diaconis, P. (1978). Review of A Mathematical Theory of Evidence. *J. Amer. Statist. Assoc.* **73**, 677–678.
- Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion (with discussion). In *Bayesian Statistics 2*, 133–156, ed. J.M. Bernardo *et al.*
- Diaconis, P. and Zabell, S. (1982). Updating subjective probability. *J. Amer. Statist. Assoc.* **77**, 822–830.
- Dias, P. and Shimony, A. (1981). A critique of Jaynes' Maximum Entropy Principle. *Advances in Applied Mathematics* **2**, 172–211.
- Dickey, J.M. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *J. Roy. Statist. Soc. Ser. B* **35**, 285–305.
- Dickey, J.M. (1974). Bayesian alternatives to the *F*-test and least-squares estimate in the normal linear model. In *Studies in Bayesian Econometrics and Statistics*, 515–554, eds. S.E. Fienberg and A. Zellner. North-Holland, Amsterdam.
- Dickey, J.M. (1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.* **71**, 680–689.
- Dickey, J.M. (1980). Beliefs about beliefs, a theory for stochastic assessments of subjective probabilities (with discussion). In *Bayesian Statistics 1*, 471–487 and 504–519, ed. J.M. Bernardo *et al.*
- Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Pr., Cambridge (MA).

- Dubins, L.E. (1975). Finitely additive conditional probabilities, conglomerability and disintegrations. *Ann. Probab.* **3**, 89–99.
- Dubins, L.E. and Savage, L.J. (1965). *How to Gamble If You Must: Inequalities for Stochastic Processes*. McGraw-Hill, New York.
- Dubois, D. and Prade, H. (1979). Decision-making under fuzziness. In *Advances in Fuzzy Set Theory and Applications*, eds. M.M. Gupta, R.K. Ragade, and R.R. Yager. North-Holland, Amsterdam.
- Dubois, D. and Prade, H. (1980). *Fuzzy Sets and Systems*. Academic Pr., New York.
- Duda, R.O. and Shortliffe, E.H. (1983). Expert systems research. *Science* **220**, 261–268.
- Dunford, N. and Schwartz, J.J. (1958). *Linear Operators, Part I: General Theory*. Interscience, New York.
- Edwards, A.W.F. (1972). *Likelihood*. Cambridge U. Pr.
- Edwards, W. (1961). Behavioral decision theory. *Ann. Rev. Psych.* **12**, 473–498.
- Edwards, W. (1968). Conservatism in human information processing. In *Formal Representation of Human Judgment*, 17–52, ed. B. Kleinmuntz. Wiley, New York.
- Edwards, W., Lindman, H. and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242.
- Eells, E. (1982). *Rational Decision and Causality*. Cambridge U. Pr.
- Einhorn, H.J. and Hogarth, R.M. (1981). Behavioral decision theory: processes of judgment and choice. *Ann. Rev. Psych.* **32**, 53–88.
- Einhorn, H.J. and Hogarth, R.M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychol. Rev.* **92**, 433–461.
- Einhorn, H.J. and Hogarth, R.M. (1986). Decision making under ambiguity. *J. Business* **59**, S225–S250.
- Ellis, R.L. (1843). On the foundations of the theory of probabilities. *Trans. Philos. Soc. Cambridge* **8**, 1–6.
- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quart. J. Econ.* **75**, 643–669.
- Elster, J. (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge U. Pr.
- Elster, J. (1983a). *Explaining Technical Change*. Cambridge U. Pr.
- Elster, J. (1983b). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge U. Pr.
- Elster, J. (ed.) (1986). *Rational Choice*. N.Y.U. Pr., New York.
- Epstein, R.A. (1977). *The Theory of Gambling and Statistical Logic*. Academic Pr., New York.
- Farquhar, P.H. (1984). Utility assessment methods. *Management Sci.* **30**, 1283–1300.

- Feagans, T.B. and Biller, W.F. (1980). Fuzzy concepts in the analysis of public health risks. In *Fuzzy Sets*, eds. P.P. Wang and S.K. Chang. Plenum Pr., New York.
- Feller, W. (1957). *Introduction to Probability Theory and its Applications*, Vol. 1 (2nd edition). Wiley, New York.
- Feller, W. (1966). *Introduction to Probability Theory and its Applications*, Vol. 2. Wiley, New York.
- Fellner, W. (1961). Distribution of subjective probabilities as a response to uncertainty. *Quart. J. Econ.* **75**, 670–689.
- Fellner, W. (1965). *Probability and Profit*. Irwin, Homewood (IL).
- Ferguson, T.S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Pr., New York.
- Fine, T.L. (1970). On the apparent convergence of relative frequency and its implications. *IEEE Trans. Inform. Theory* **IT-16**, 251–257.
- Fine, T.L. (1973). *Theories of Probability*. Academic Pr., New York.
- Fine, T.L. (1977a). An argument for comparative probability. In *Basic Problems in Methodology and Linguistics*, 105–119, eds. R.E. Butts and J. Hintikka. Reidel, Dordrecht.
- Fine, T.L. (1977b). Review of A Mathematical Theory of Evidence. *Bull. Amer. Math. Soc.* **83**, 667–672.
- Fine, T.L. (1978). Review of The Emergence of Probability. *Philos. Rev.* **87**, 116–123.
- Fine, T.L. (1983). Foundations of probability. In *The Encyclopedia of Statistical Sciences*, Vol. 3, eds. S. Kotz and N.L. Johnson. Wiley, New York.
- Fine, T.L. (1988). Lower probability models for uncertainty and nondeterministic processes. *J. Statist. Planning Inf.* **20**, 389–411.
- de Finetti, B. (1931). Sul significato soggettivo della probabilità. *Fund. Math.* **17**, 298–329.
- de Finetti, B. (1937). La prevision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68. English translation in Kyburg and Smokler (1964).
- de Finetti, B. (1972). *Probability, Induction and Statistics*. Wiley, London.
- de Finetti, B. (1974). *Theory of Probability*, Vol. 1. English translation of *Teoria delle Probabilità* (1970). Wiley, London.
- de Finetti, B. (1975). *Theory of Probability*, Vol. 2. Wiley, London.
- de Finetti, B. (1977a). Probabilities of probabilities: a real problem or a misunderstanding? In *New Developments in the Applications of Bayesian Methods*, 1–10, eds. A. Aykac and C. Brumat. North-Holland, Amsterdam.
- de Finetti, B. (1977b). Probability: Beware of falsifications! In *New*

- Developments in the Applications of Bayesian Methods*, 347–385, eds. A. Aykac and C. Brumat. North-Holland, Amsterdam.
- de Finetti, B. and Savage, L.J. (1962). Sul modo di scegliere le probabilità iniziali. *Biblioteca del Metron*, Ser. C, Vol. I, 81–147. English summary in de Finetti (1972) and in Savage (1981).
- Fishburn, P.C. (1964). *Decision and Value Theory*. Wiley, New York.
- Fishburn, P.C. (1965). Analysis of decisions with incomplete knowledge of probabilities. *Operations Research* **13**, 217–237.
- Fishburn, P.C. (1970). *Utility Theory for Decision Making*. Wiley, New York.
- Fishburn, P.C. (1975). A theory of subjective expected utility with vague preferences. *Theory and Decision* **6**, 287–310.
- Fishburn, P.C. (1981). Subjective expected utility: a review of normative theories. *Theory and Decision* **13**, 139–199.
- Fishburn, P.C. (1986). The axioms of subjective probability (with discussion). *Statist. Sci.* **1**, 335–358.
- Fishburn, P.C., Murphy, A.H. and Isaacs, H.H. (1968). Sensitivity of decisions to probability estimation errors: a re-examination. *Operations Research* **16**, 253–268.
- Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Camb. Philos. Soc.* **22**, 700–725.
- Fisher, R.A. (1935). The fiducial argument in statistical inference. *Ann. Eugenics London* **6**, 391–398.
- Fisher, R.A. (1956a). *Statistical Methods and Scientific Inference* (3rd edition, 1973). Oliver and Boyd, Edinburgh.
- Fisher, R.A. (1956b). On a test of significance in Pearson's Biometrika tables. *J. Roy. Statist. Soc. Ser. B* **18**, 56–60.
- Fisher, R.A. (1957). The underworld of probability. *Sankhyā* **18**, 201–210.
- Fisher, R.A. (1958). The nature of probability. *Centennial Review* **2**, 261–274.
- Fodor, J.A. (1975). *The Language of Thought*. Crowell, New York.
- Forsyth, R. (ed.) (1984). *Expert Systems*. Chapman and Hall, London.
- Fraser, D.A.S. (1961a). On fiducial inference. *Ann. Math. Statist.* **32**, 661–676.
- Fraser, D.A.S. (1961b). The fiducial method and invariance. *Biometrika* **48**, 261–280.
- Fraser, D.A.S. (1968). *The Structure of Inference*. Wiley, New York.
- Fraser, D.A.S. (1972). Bayes, likelihood, or structural. *Ann. Math. Statist.* **43**, 777–790.
- Fraser, D.A.S. (1977). Confidence, posterior probability, and the Buehler example. *Ann. Statist.* **5**, 892–898.
- Fraser, D.A.S. (1979). *Inference and Linear Models*. McGraw-Hill, New York.
- Freedman, D.A. and Purves, R.A. (1969). Bayes' method for bookies. *Ann. Math. Statist.* **40**, 1177–1186.
- Freeling, A.N.S. (1980). Fuzzy sets and decision analysis. *IEEE Trans. SMC-10*, 341–354.

- French, S. (1982). On the axiomatisation of subjective probabilities. *Theory and Decision* **14**, 19–33.
- Gaines, B.R. (1975). Stochastic and fuzzy logics. *Electronics Letters* **11**, 188–189.
- Gaines, B.R. (1976). Foundations of fuzzy reasoning. *Internat. J. Man-Machine Studies* **8**, 623–668.
- Gaines, B.R. (1978). Fuzzy and probability uncertainty logics. *Inform. Control* **38**, 154–169.
- Gale, D. (1960). *The Theory of Linear Economic Models*. McGraw-Hill, New York.
- Gale, W.A. (ed.) (1986). *Artificial Intelligence and Statistics*. Addison-Wesley, Reading (MA).
- Gardenfors, P. and Sahlin, N. (1982). Unreliable probabilities, risk-taking, and decision making. *Synthese* **53**, 361–386.
- Gardenfors, P. and Sahlin, N. (1983). Decision making with unreliable probabilities. *Brit. J. Math. Statist. Psychol.* **36**, 240–251.
- Gardner, H. (1985). *The Mind's New Science*. Basic Books, New York.
- Gauss, C.F. (1809). *Theoria Motus Corporum Celestium in Sectionibus Conicis Solum Ambientium*. English translation by C.H. Davis (1963), Dover, New York.
- Geisser, S. (1984). On prior distributions for binary trials. *Amer. Statist.* **38**, 244–251.
- Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*. Huebner Foundation, Philadelphia.
- Giere, R.N. (1973). Objective single-case probabilities and the foundations of statistics. In *Logic, Methodology and Philosophy of Science IV*, 467–483, eds. P. Suppes et al. North-Holland, Amsterdam.
- Giere, R.N. (1976). Empirical probability, objective statistical methods, and scientific inquiry. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol. II, 63–101, eds. W.L. Harper and C.A. Hooker. Reidel, Dordrecht.
- Giles, R. (1976). A logic for subjective belief. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Vol. I, 41–70, eds. W. L. Harper and C. A. Hooker. Reidel, Dordrecht.
- Giles, R. (1988). The concept of grade of membership. *Fuzzy Sets and Systems* **25**, 297–323.
- Gillies, D.A. (1973). *An Objective Theory of Probability*. Methuen, London.
- Giron, F.J. and Rios, S. (1980). Quasi-Bayesian behaviour: a more realistic approach to decision making? (with discussion). In *Bayesian Statistics 1*, 17–38 and 49–66, eds. J.M. Bernardo et al.
- Goel, P.K. and DeGroot, M.H. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.* **76**, 140–147.
- Goldstein, M. (1974). Approximate Bayesian inference with incompletely specified prior distributions. *Biometrika* **61**, 629–631.

- Goldstein, M. (1981). Revising previsions: a geometric interpretation (with discussion). *J. Roy. Statist. Soc. Ser. B* **43**, 105–130.
- Goldstein, M. (1983). The prevision of a prevision. *J. Amer. Statist. Assoc.* **78**, 817–819.
- Goldstein, M. (1985). Temporal coherence (with discussion). In *Bayesian Statistics 2*, 231–248, ed. J.M. Bernardo *et al.*
- Goldstein, M. (1986). Separating beliefs. In *Bayesian Inference and Decision Techniques*, eds. P.K. Goel and A. Zellner. North-Holland, Amsterdam.
- Good, I.J. (1950). *Probability and the Weighing of Evidence*. Griffin, London.
- Good, I.J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B* **14**, 107–114. Reprinted in Good (1983).
- Good, I.J. (1962a). Subjective probability as the measure of a non-measurable set. In *Logic, Methodology and Philosophy of Science*, 319–329, eds. E. Nagel, P. Suppes, and A. Tarski. Stanford U. Pr. Reprinted in Good (1983).
- Good, I.J. (1962b). How rational should a manager be? *Management Sci.* **8**, 383–393.
- Good, I.J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34**, 911–934.
- Good, I.J. (1965). *The Estimation of Probabilities*. MIT Pr., Cambridge (MA).
- Good, I.J. (1966). A derivation of the probabilistic explication of information. *J. Roy. Statist. Soc. Ser. B* **28**, 578–581.
- Good, I.J. (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. Ser. B* **29**, 399–431.
- Good, I.J. (1971a). The probabilistic explication of information, evidence, surprise, causality, explanation, and utility. In *Foundations of Statistical Inference*, 108–141, eds. V.P. Godambe and D.A. Sprott. Holt, Rinehart, and Winston, Toronto. Reprinted in Good (1983).
- Good, I.J. (1971b). 46656 varieties of Bayesians. *Amer. Statistician* **25**, 62–63. Reprinted in Good (1983).
- Good, I.J. (1975). Explicativity, corroboration, and the relative odds of hypotheses. *Synthese* **30**, 39–73. Reprinted in Good (1983).
- Good, I.J. (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. II, eds. W.L. Harper and C.A. Hooker. Reidel, Dordrecht. Reprinted in Good (1983).
- Good, I.J. (1980). Some history of the hierarchical Bayesian methodology (with discussion). In *Bayesian Statistics 1*, 489–519, ed. J.M. Bernardo *et al.* Reprinted in Good (1983).
- Good, I.J. (1983). *Good Thinking*. U. Minnesota Pr., Minneapolis.
- Good, I.J. (1985). Weight of evidence: a brief survey (with discussion). In *Bayesian Statistics 2*, 249–269, ed. J.M. Bernardo *et al.*

- Granirer, E. (1973). Criteria for compactness and for discreteness of locally compact amenable groups. *Proc. Amer. Math. Soc.* **40**, 615–624.
- Greenleaf, F.P. (1969). *Invariant Means on Topological Groups and Their Applications*. Van Nostrand, New York.
- Griffiths, A.P. (ed.) (1967). *Knowledge and Belief*. Oxford U. Pr.
- Grize, Y.-L. and Fine, T.L. (1987). Continuous lower probability-based models for stationary processes with bounded and divergent time averages. *Ann. Probab.* **15**, 783–803.
- Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge U. Pr.
- Hacking, I. (1967). Slightly more realistic personal probability. *Philos. Sci.* **34**, 311–325.
- Hacking, I. (1974). Combined evidence. In *Logical Theory and Semantic Analysis*, ed. S. Stenlund. Reidel, Dordrecht.
- Hacking, I. (1975). *The Emergence of Probability*. Cambridge U. Pr.
- Hailperin, T. (1976). *Boole's Logic and Probability*. North-Holland, Amsterdam.
- Haldane, J.B.S. (1945). On a method of estimating frequencies. *Biometrika* **33**, 222–225.
- Halmos, P.R. (1950). *Measure Theory*. Van Nostrand, New York.
- Hamblin, C.L. (1959). The modal ‘probably’. *Mind* **68**, 234–240.
- Hampton, J.M., Moore, P.G. and Thomas, H. (1973). Subjective probability and its measurement. *J. Roy. Statist. Soc. Ser. A* **136**, 21–42.
- Harsanyi, J.C. (1977). *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge U. Pr.
- Hartigan, J.A. (1964). Invariant prior distributions. *Ann. Math. Statist.* **35**, 836–845.
- Hartigan, J.A. (1969). Linear Bayesian methods. *J. Roy. Statist. Soc. Ser. B* **31**, 446–454.
- Hartigan, J.A. (1983). *Bayes Theory*. Springer-Verlag, New York.
- Heath, D.C. and Sudderth, W.D. (1972). On a theorem of de Finetti, oddsmaking, and game theory. *Ann. Math. Statist.* **6**, 2072–2077.
- Heath, D.C. and Sudderth, W.D. (1976). De Finetti's theorem on exchangeable variables. *Amer. Statist.* **30**, 188–189.
- Heath, D.C. and Sudderth, W.D. (1978). On finitely additive priors, coherence, and extended admissibility. *Ann. Statist.* **6**, 333–345.
- Hempel, C.G. (1965). *Aspects of Scientific Explanation*. Free Pr., New York.
- Hempel, C.G. (1966). *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs (NJ).
- Hersh, H.M. and Caramazza, A. (1976). A fuzzy set approach to modifiers and vagueness in natural language. *J. Exper. Psychology* **105**, 254–276.
- Hewitt, E. and Savage, L.J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80**, 470–501.

- Hey, J.D. (1979). *Uncertainty in Microeconomics*. New York U. Pr.
- 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*, 555–584, eds. S.E. Fienberg and A. Zellner. North-Holland, Amsterdam.
- Hill, B.M. (1980). On some statistical paradoxes and non-conglomerability (with discussion). In *Bayesian Statistics 1*, 39–66, eds. J.M. Bernardo *et al.*
- Hill, B.M. and Lane, D.A. (1986). Conglomerability and countable additivity. In *Bayesian Inference and Decision Techniques*, eds. P.K. Goel and A. Zellner. North-Holland, Amsterdam.
- Hisdal, E. (1988). Are grades of membership probabilities? *Fuzzy Sets and Systems* 25, 325–348.
- Hodges, J.L. and Lehmann, E.L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Statist.* 23, 396–407.
- Hogarth, R.M. (1975). Cognitive processes and the assessment of subjective probability distributions. *J. Amer. Statist. Assoc.* 70, 271–294.
- Hogarth, R.M. (1980). *Judgement and Choice*. Wiley, New York.
- Hogarth, R.M. and Kunreuther, H. (1985). Ambiguity and insurance decisions. *Amer. Economic Rev.* 75, 386–390.
- Holmes, R.B. (1975). *Geometric Functional Analysis and its Applications*. Springer-Verlag, New York.
- Hora, R.B. and Buehler, R.J. (1966). Fiducial theory and invariant estimation. *Ann. Math. Statist.* 37, 643–656.
- Huber, G.P. (1974). Methods for quantifying subjective probabilities and multi-attribute utilities. *Decision Sci.* 5, 430–458.
- Huber, P.J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* 36, 1753–1758.
- Huber, P.J. (1973). The use of Choquet capacities in statistics. *Bull. Internat. Statist. Inst.* 45, Book 4, 181–188.
- Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.
- Huber, P.J. and Strassen, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* 1, 251–263.
- Hull, C.L. (1943). *Principles of Behavior*. Appleton-Century-Crofts, New York.
- Isaacs, H.H. (1963). Sensitivity of decisions to probability estimation errors. *Operations Research* 11, 536–552.
- Jain, R. (1976). Decision making in the presence of fuzzy variables. *IEEE Trans. SMC-6*, 698–703.
- Jameson, G.J.O. (1970). *Ordered Linear Spaces*. Springer-Verlag, Berlin.
- Jameson, G.J.O. (1974). *Topology and Normed Spaces*. Chapman and Hall, London.
- Janis, I.L. and Mann, L. (1977). *Decision Making*. Free Pr., New York.

- Jaynes, E.T. (1957). Information theory and statistical mechanics. *Phys. Rev.* 106, 620–630, and 108, 171–190. Reprinted in Jaynes (1983).
- Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. Systems Science and Cybernetics*, SSC-4, 227–241. Reprinted in Jaynes (1983).
- Jaynes, E.T. (1980). Marginalization and prior probabilities. In *Bayesian Analysis in Econometrics and Statistics*, ed. A. Zellner. North-Holland, Amsterdam. Reprinted in Jaynes (1983).
- Jaynes, E.T. (1983). *Papers on Probability, Statistics and Statistical Physics*, ed. R.D. Rosenkrantz. Reidel, Dordrecht.
- Jaynes, E.T. (1986). Some applications and extensions of the de Finetti representation theorem. In *Bayesian Inference and Decision Techniques*, eds. P.K. Goel and A. Zellner. North-Holland, Amsterdam.
- Jech, T.J. (1973). *The Axiom of Choice*. North-Holland, Amsterdam.
- Jeffrey, R.C. (1968). Probable knowledge. In *The Problem of Inductive Logic*, ed. I. Lakatos. North-Holland, Amsterdam.
- Jeffrey, R.C. (1983). *The Logic of Decision*, 2nd edition. U. Chicago Pr.
- Jeffreys, H. (1931). *Scientific Inference*. Cambridge U. Pr.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. Ser. A* 186, 453–461.
- Jeffreys, H. (1983). *Theory of Probability*, 3rd edition. (1st edition 1939). Clarendon Pr., Oxford.
- Johnson, W.E. (1924). *Logic*, Part III. Cambridge U. Pr.
- Johnson, W.E. (1932). Appendix to ‘Probability: deductive and inductive problems.’ *Mind* 41, 421–423.
- Kadane, J.B. and Chuang, D.T. (1978). Stable decision problems. *Ann. Statist.* 6, 1095–1110.
- Kadane, J.B., Schervish, M.J. and Seidenfeld, T. (1986). Statistical implications of finitely additive probability. In *Bayesian Inference and Decision Techniques*, eds. P.K. Goel and A. Zellner. North-Holland, Amsterdam.
- Kahneman, D., Slovic, P. and Tversky, A. (eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge U. Pr.
- Kanal, L.N. and Lemmer, J.F. (eds.) (1986). *Uncertainty in Artificial Intelligence*. North-Holland, Amsterdam.
- Kaplan, M. and Fine, T.L. (1977). Joint orders in comparative probability. *Ann. Probab.* 5, 161–179.
- Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. *Combinatorica* 4, 373–395.
- Kekes, J. (1976). *A Justification of Rationality*. SUNY Pr., Albany (NY).
- Kelley, J.L. (1955). *General Topology*. Van Nostrand, New York.
- Kelley, J.L. and Namioka, I. (1963). *Linear Topological Spaces*. Van Nostrand, Princeton (NJ).

- Kemeny, J. (1955). Fair bets and inductive probabilities. *J. Symbolic Logic* **20**, 263–273.
- Kempthorne, O. (1966). Some aspects of experimental inference. *J. Amer. Statist. Assoc.* **61**, 11–34.
- Kempthorne, O. and Folks, L. (1971). *Probability, Statistics and Data Analysis*. Iowa State U. Pr., Ames (IA).
- Kent, S. (1964). Words of estimated probability. *Studies Intelligence* **8**, 49–65.
- Keynes, J.M. (1921). *A Treatise on Probability*. Vol. 8 of Collected Writings (1973 edition). Macmillan, London.
- Kickert, W.J.M. (1978). *Fuzzy Theories on Decision-Making*. Nijhoff, Leiden.
- Kiefer, J. (1977). Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72**, 789–827.
- Kingman, J.F.C. (1978). The uses of exchangeability. *Ann. Probab.* **16**, 183–197.
- Kmietowicz, Z.W. and Pearman, A.D. (1981). *Decision Theory and Incomplete Knowledge*. Gower, Aldershot (UK).
- Kneale, W. (1949). *Probability and Induction*. Oxford U. Pr., Oxford.
- Knight, F.H. (1933). *Risk, Uncertainty and Profit*. Houghton Mifflin, Boston. (First published 1921.)
- Kolmogorov, A.N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*. Second English edition (1956), *Foundations of the Theory of Probability*. Chelsea, New York.
- Koopman, B.O. (1940a). The bases of probability. *Bull. Amer. Math. Soc.* **46**, 763–774. Reprinted in Kyburg and Smokler (1964).
- Koopman, B.O. (1940b). The axioms and algebra of intuitive probability. *Ann. Math.* **41**, 269–292.
- Koopman, B.O. (1941). Intuitive probabilities and sequences. *Ann. Math.* **42**, 169–187.
- Kraft, C., Pratt, J. and Seidenberg, A. (1959). Intuitive probability on finite sets. *Ann. Math. Statist.* **30**, 408–419.
- Krantz, D.H., Luce, R.D., Suppes, P. and Tversky, A. (1971). *Foundations of Measurement*, Vol. 1. Academic Pr., New York.
- Kudō, H. (1967). On partial prior information and the property of parametric sufficiency. In *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1**, 251–265. U. California Pr., Berkeley.
- Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. U. Chicago Pr.
- Kuhn, T.S. (1963). The function of dogma in scientific research (with discussion). In *Scientific Change*, 347–369, ed. A.C. Crombie. Basic Books, New York.
- Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.
- Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

- Kumar, A. and Fine, T.L. (1985). Stationary lower probabilities and unstable averages. *Z. Wahr. v. Gebiete* **69**, 1–17.
- Kyburg, H.E. (1961). *Probability and the Logic of Rational Belief*. Wesleyan U. Pr., Middletown (CT).
- Kyburg, H.E. (1970). *Probability and Inductive Logic*. Macmillan, New York.
- Kyburg, H.E. (1974a). *The Logical Foundations of Statistical Inference*. Reidel, Dordrecht.
- Kyburg, H.E. (1974b). Propensities and probabilities. *Brit. J. Philos. Sci.* **25**, 358–375. Reprinted in Tuomela (1978).
- Kyburg, H.E. (1978). Subjective probability: criticisms, reflections, and problems. *J. Phil. Logic* **7**, 157–180.
- Kyburg, H.E. (1983a). *Epistemology and Inference*. U. Minnesota Pr., Minneapolis.
- Kyburg, H.E. (1983b). Rational belief (with discussion). *Behav. Brain Sci.* **6**, 231–273.
- Kyburg, H.E. and Smokler, H.E. (eds.) (1964). *Studies in Subjective Probability*. Wiley, New York. Second edition (with new material) 1980.
- Lad, F.R. (1990). *Operational Subjective Statistical Methods: A Mathematical, Philosophical and Historical Introduction, with Applications*. (Draft of Chapter 1.) U. Canterbury, Christchurch (NZ).
- Lane, D.A. and Sudderth, W.D. (1983). Coherent and continuous inference. *Ann. Statist.* **11**, 114–120.
- Laplace, P.S. de (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.
- Laplace, P.S. de (1814). *Essai Philosophique sur les Probabilités*. English translation by Truscott and Emory (1952). Dover, New York.
- Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50**, 157–224.
- Leamer, E.E. (1978). *Specification Searches*. Wiley, New York.
- Leamer, E.E. (1982). Sets of posterior means with bounded variance prior. *Econometrica* **50**, 725–736.
- Leamer, E.E. (1986). Bid-ask spreads for subjective probabilities. In *Bayesian Inference and Decision Techniques*, eds. P.K. Goel and A. Zellner. North-Holland, Amsterdam.
- Lehman, R.S. (1955). On confirmation and rational betting. *J. Symbolic Logic* **20**, 251–262.
- Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd edition. Wiley, New York.
- Levi, I. (1967). *Gambling with Truth*. Knopf, New York.
- Levi, I. (1974). On indeterminate probabilities. *J. Philosophy* **71**, 391–418.
- Levi, I. (1980). *The Enterprise of Knowledge*. MIT Pr., London.
- Levi, I. (1982). Ignorance, probability and rational choice. *Synthese* **53**, 387–417.

- Levi, I. (1985). Imprecision and indeterminacy in probability judgment. *Philos. Sci.* **52**, 390–409.
- Levine, R.D. and Tribus, M. (eds.) (1979). *The Maximum Entropy Formalism*. MIT Pr., Cambridge (MA).
- Lewis, D. (1980). A subjectivist's guide to objective chance. In *Studies in Inductive Logic and Probability*, Vol. II, 263–293, ed. R.C. Jeffrey. U. California Pr., Berkeley.
- Lichtenstein, S., Fischhoff, B. and Phillips, L.D. (1977). Calibration of probabilities: the state of the art. In *Decision Making and Change in Human Affairs*, 275–324, eds. H. Jungerman and G. de Zeeuw. Reidel, Dordrecht.
- Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.
- Lindley, D.V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.
- Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem. *J. Roy. Statist. Soc. Ser. B* **20**, 102–107.
- Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*. Cambridge U. Pr.
- Lindley, D.V. (1971a). *Making Decisions*. Wiley, London.
- Lindley, D.V. (1971b). *Bayesian Statistics, A Review*. Society for Industrial and Applied Mathematics, Philadelphia.
- Lindley, D.V. (1982). Scoring rules and the inevitability of probability (with discussion). *Internat. Statist. Rev.* **50**, 1–26.
- Lindley, D.V. (1983). Theory and practice of Bayesian statistics. *Statistician* **32**, 1–11.
- Lindley, D.V. and Novick, M.R. (1981). The role of exchangeability in inference. *Ann. Statist.* **9**, 45–58.
- Lindley, D.V. and Phillips, L.D. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.* **30**, 112–119.
- Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 1–41.
- Lindley, D.V., Tversky, A. and Brown, R.V. (1979). On the reconciliation of probability assessments (with discussion). *J. Roy. Statist. Soc. Ser. A* **142**, 146–180.
- Loomis, L.H. (1953). *An Introduction to Abstract Harmonic Analysis*. Van Nostrand, Princeton (NJ).
- Machlup, F. and Mansfield, U. (eds.) (1983). *The Study of Information*. Wiley, New York.
- Mackenzie, B.D. (1977). *Behaviourism and the Limits of Scientific Method*. Humanities Pr., Atlantic Highlands (NJ).
- Mallows, C.L. and Walley, P. (1980). A theory of data analysis? *Bell Laboratories Technical Memorandum TM 80-1215-7*, Murray Hill (NJ).

- Manski, C.F. (1981). Learning and decision making when subjective probabilities have subjective domains. *Ann. Statist.* **9**, 59–65.
- March, J.G. (1978). Bounded rationality, ambiguity, and the engineering of choice. *Bell J. Econ.* **9**, 587–608.
- McClure, J. (1984). *Development and Applications of the Calculus of Interval-Valued Probability*. Ph.D. Dissertation, Cornell U., Ithaca (NY).
- Mellor, D.H. (1967). Imprecision and explanation. *Philos. Sci.* **34**, 1–9.
- Mellor, D.H. (1971). *The Matter of Chance*. Cambridge U. Pr.
- Mellor, D.H. (1974). In defense of dispositions. *Philos. Rev.* **83**, 157–181. Reprinted in Tuomela (1978).
- Mellor, D.H. (1980). Consciousness and degrees of belief. In *Prospects for Pragmatism*, ed. D.H. Mellor. Cambridge U. Pr.
- Menges, G. (1966). On the Bayesification of the minimax principle. *Unternehmensforschung* **10**, 81–91.
- Mill, J.S. (1843). *A System of Logic: Ratiocinative and Inductive*. Parker, London.
- von Mises, R. (1942). On the correct use of Bayes' formula. *Ann. Math. Statist.* **13**, 156–165.
- von Mises, R. (1957). *Probability, Statistics and Truth*. (English translation of 3rd German edition, 1951.) George Allen & Unwin, London.
- Mood, A.M., Graybill, F.A. and Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd edition. McGraw-Hill, New York.
- Moore, G.H. (1982). *Zermelo's Axiom of Choice*. Springer-Verlag, New York.
- Mosteller, F. (1965). *Fifty Challenging Problems in Probability*. Addison-Wesley, Reading (MA).
- Mosteller, F. and Wallace, D.L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading (MA).
- Nachbin, L. (1965). *The Haar Integral*. Van Nostrand, Princeton (NJ).
- Nagel, E. (1961). *The Structure of Science*. Harcourt, Brace & World, New York.
- Nathanson, S. (1985). *The Ideal of Rationality*. Humanities, Atlantic Highlands (NJ).
- Nau, R.F. (1981). Coherent assessment of subjective probability. *ORC 81-5*, Operations Research Center, U. California, Berkeley.
- Nau, R.F. (1986). A new theory of indeterminate probabilities and utilities. *Working paper No. 8609*, Fuqua School of Business, Duke U.
- von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd edition. Princeton U. Pr.
- Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Roy. Soc. Ser. A* **236**, 333–380.
- Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika* **32**, 128–150.

- Neyman, J. (1957). 'Inductive behavior' as a basic concept of philosophy of science. *Rev. Internat. Statist. Inst.* **25**, 7–22.
- Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36**, 97–131.
- Nisbett, R. and Ross, L. (1980). *Human Inference*. Prentice-Hall, Englewood Cliffs (NJ).
- Nisbett, R.E., Krantz, D.H., Jepson, C. and Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychol. Rev.* **90**, 339–363.
- Norman, R. (1971). *Reasons for Action*. Blackwell, Oxford.
- Novick, M.R. (1969). Multiparameter Bayesian indifference procedures (with discussion). *J. Roy. Statist. Soc. Ser. B* **31**, 29–64.
- Novick, M.R. and Hall, W.J. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**, 1104–1117.
- Olshen, R.A. (1973). The conditional level of the *F*-test. *J. Amer. Statist. Assoc.* **68**, 692–698.
- Papamarcou, A. (1987). *Unstable Random Sequences as an Objective Basis for Interval-Valued Probability Models*. Ph.D. Dissertation, Cornell U., Ithaca (NY).
- Papamarcou, A. and Fine, T.L. (1986). A note on undominated lower probabilities. *Ann. Probab.* **14**, 710–723.
- Parikh, R. and Parnes, M. (1974). Conditional probabilities and uniform sets. In *Victoria Symposium on Nonstandard Analysis*, 180–194, eds. A. Hurd and P. Loeb. Lecture Notes in Math. 369. Springer-Verlag, Berlin.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo (CA).
- Pears, D.F. (1984). *Motivated Irrationality*. Clarendon Pr., Oxford.
- Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika* **13**, 1–16.
- Pearson, K. (1921). Note on the 'fundamental problem of practical statistics'. *Biometrika* **13**, 300–301.
- Pedersen, J.G. (1978). Fiducial inference. *Internat. Statist. Rev.* **46**, 147–170.
- Peirce, C.S. (1878). The probability of induction. *Popular Science Monthly*. Reprinted in *Philosophical Writings of Peirce*, Ch. 13, ed. J. Buchler (1955). Dover, New York.
- Pereira, C.A.B. and Pericchi, L.R. (1988). Analysis of diagnosability. Tech. Report, U. São Paulo. To appear in *Applied Statistics* **39**, no. 1.
- Pericchi, L.R. and Nazaret, W.A. (1987). On being imprecise at the higher levels of a hierarchical linear model. To appear in *Bayesian Statistics 3*, eds. J.M. Bernardo *et al.*
- Pericchi, L.R. and Walley, P. (1989). One-sided hypothesis testing with near-ignorance priors. To appear in *Rev. Brasil. Probab. Estat. (REBRAPE)*.

- Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *J. Inst. Actuar.* **73**, 285–312.
- Pierce, D.A. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.* **1**, 241–250.
- Pierce, D.A. and Folks, J.L. (1969). Sensitivity of Bayes procedures to the prior distribution. *Operations Research* **17**, 344–350.
- Pincus, D. (1974). The strength of the Hahn-Banach theorem. In *Victoria Symposium on Nonstandard Analysis*, 203–248, eds. A. Hurd and P. Loeb. Lecture Notes in Mathematics 369. Springer-Verlag, Berlin.
- Pitz, G.F. and Sachs, N.J. (1984). Judgment and decision: theory and application. *Ann. Rev. Psychol.* **35**, 139–163.
- Polasek, W. (1984). Multivariate regression systems: estimation and sensitivity analysis of two-dimensional data. In *Robustness of Bayesian Analyses*, ed. J.B. Kadane. Elsevier Science, Amsterdam.
- Polasek, W. (1986). Local sensitivity analysis and Bayesian regression diagnostics. In *Bayesian Inference and Decision Techniques*, eds. P.K. Goel and A. Zellner. North-Holland, Amsterdam.
- Popper, K.R. (1957). Probability magic, or knowing out of ignorance. *Dialectica* **11**, 354–373.
- Popper, K.R. (1959a). *The Logic of Scientific Discovery*. Hutchinson, London.
- Popper, K.R. (1959b). The propensity interpretation of probability. *Brit. J. Philos. Sci.* **10**, 25–42. Reprinted in Tuomela (1978).
- Popper, K.R. (1962). *Conjectures and Refutations*. Basic Books, New York.
- Popper, K.R. (1983). *Realism and the Aim of Science*. Hutchinson, London.
- Potter, J.M. and Anderson, B.D.O. (1980). Partial prior information and decision making. *IEEE Trans. SMC-10*, 125–133.
- Potter, J.M. and Anderson, B.D.O. (1983). Statistical inference with partial prior information. *IEEE Trans. IT-29*, 688–695.
- Prade, H. (1985). A computational approach to approximate and plausible reasoning with applications to expert systems. *IEEE Trans. PAMI-7*, 260–283.
- Pratt, J.W. (1961). Review of Testing Statistical Hypotheses. *J. Amer. Statist. Assoc.* **56**, 163–167.
- Pratt, J.W., Raiffa, H. and Schlaifer, R. (1964). The foundations of decision under uncertainty: an elementary exposition. *J. Amer. Statist. Assoc.* **59**, 353–375.
- Putnam, H. (1981). *Reason, Truth, and History*. Cambridge U. Pr.
- Quinlan, J.R. (1983). Inferno: a cautious approach to uncertain inference. *Computer J.* **26**, 255–269.
- Raiffa, H. (1968). *Decision Analysis*. Addison-Wesley, Reading (MA).
- Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard U. Pr.

- Ramsey, F.P. (1926). Truth and probability. In Ramsey (1931), 156–198, and in Kyburg and Smokler (1964), 61–92.
- Ramsey, F.P. (1931). *The Foundations of Mathematics and other Logical Essays*, ed. R.B. Braithwaite. Kegan Paul, Trench, Trubner; London.
- Rawls, J. (1971). *A Theory of Justice*. Harvard U. Pr., Cambridge (MA).
- Raz, J. (ed.) (1978). *Practical Reasoning*. Oxford U. Pr.
- Regazzini, E. (1983). Coherent conditional probabilities, finite additivity, extensions. *Tech. Report*. Dipartimento di Scienze Statistiche, U. Bologna.
- Regazzini, E. (1987). De Finetti's coherence and statistical inference. *Ann. Statist.* **15**, 845–864.
- Reichenbach, H. (1949). *The Theory of Probability*. U. California Pr., Berkeley.
- Rényi, A. (1955). On a new axiomatic theory of probability. *Acta Math. Acad. Sci. Hung.* **6**, 285–335.
- Rescher, N. (1968). *Topics in Philosophical Logic*. Reidel, Dordrecht.
- Robinson, G.K. (1975). Some counterexamples to the theory of confidence intervals. *Biometrika* **62**, 155–161.
- Robinson, G.K. (1976). Properties of Student's *t* and of the Behrens–Fisher solution to the two means problem. *Ann. Statist.* **4**, 963–971.
- Robinson, G.K. (1977). Conservative statistical inference. *J. Roy. Statist. Soc. Ser. B* **39**, 381–386.
- Robinson, G.K. (1979a). Conditional properties of statistical procedures. *Ann. Statist.* **7**, 742–755.
- Robinson, G.K. (1979b). Conditional properties of statistical procedures for location and scale parameters. *Ann. Statist.* **7**, 756–771.
- Robinson, G.K. (1982). Behrens–Fisher problem. In *Encyclopedia of Statistical Sciences*, Vol 1. Wiley, New York.
- Rockafellar, R.T. (1970). *Convex Analysis*. Princeton U. Pr.
- Rorty, R. (1980). *Philosophy and the Mirror of Nature*. Princeton U. Pr.
- Rosenkrantz, R.D. (1977). *Inference, Method, and Decision*. Reidel, Boston.
- Royden, H.L. (1963). *Real Analysis*. Macmillan, New York.
- Rubin, H. (1985). A weak system of axioms for “rational” behavior and the non-separability of utility from prior. *Technical Report*, Dept. of Statistics, Purdue U.
- Russell, B. (1948). *Human Knowledge: Its Scope and Limits*. George Allen and Unwin, London.
- Ryle, G. (1949). *The Concept of Mind*. Hutchinson, London.
- Salmon, W.C. (1966). *The Foundations of Scientific Inference*. U. Pittsburgh Pr.
- Savage, L.J. (1967). Difficulties in the theory of personal probability. *Philos. Sci.* **34**, 305–310. Reprinted in Savage (1981).

- Savage, L.J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**, 783–801. Reprinted in Savage (1981).
- Savage, L.J. (1972a). *The Foundations of Statistics*, 2nd revised edition (first published 1954). Dover, New York.
- Savage, L.J. (1972b). Diagnosis and the Bayesian viewpoint. In *Computer Diagnosis and Diagnostic Methods*, 131–138, ed. J.A. Jacquez. Thomas, Springfield (IL). Reprinted in Savage (1981).
- Savage, L.J. (1977). The shifting foundations of statistics. In *Logic, Laws and Life*, 3–18, ed. R. Colodny. U. Pittsburgh Pr. Reprinted in Savage (1981).
- Savage, L.J. (1981). *The Writings of Leonard Jimmie Savage – A Memorial Selection*. ASA & IMS, Washington D.C.
- Schefe, P. (1980). On foundations of reasoning with uncertain facts and vague concepts. *Internat. J. Man-Machine Studies* **12**, 35–62.
- Schervish, M.J., Seidenfeld, T. and Kadane, J.B. (1984). The extent of non-conglomerability of finitely additive probabilities. *Z. Wahr. v. Gebiete* **66**, 205–226.
- Schoemaker, P.J.H. (1982). The expected utility model: its variants, purposes, evidence and limitations. *J. Econ. Literature* **20**, 529–563.
- Schum, D.A. (1979). A review of a case against Blaise Pascal and his heirs. *U. Michigan Law Rev.* **77**, 446–483.
- Schum, D.A. and Martin, A.W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Rev.* **17**, 105–157.
- Scott, D. (1964). Measurement structures and linear inequalities. *J. Math. Psychol.* **1**, 233–247.
- Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference*. Reidel, Dordrecht.
- Sen, A.K. (1982). *Choice, Welfare and Measurement*. MIT Pr., Cambridge (MA).
- Shackle, G.L.S. (1952). *Expectation in Economics*, 2nd edition. Cambridge U. Pr.
- Shackle, G.L.S. (1961). *Decision, Order and Time in Human Affairs*. Cambridge U. Pr.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton U. Pr.
- Shafer, G. (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences* **19**, 309–370.
- Shafer, G. (1979). Allocations of probability. *Ann. Probab.* **7**, 827–839.
- Shafer, G. (1981a). Constructive probability. *Synthese* **48**, 1–60.
- Shafer, G. (1981b). Two theories of probability. In *Philosophy of Science Association Proceedings* 1978, Vol. 2, eds. P. Asquith and I. Hacking. Philosophy of Science Association, East Lansing (MI).

- Shafer, G. (1981c). Jeffrey's rule of conditioning. *Philos. Sci.* **48**, 337–362.
- Shafer, G. (1981d). When are conditional probabilities legitimate? Unpublished manuscript.
- Shafer, G. (1982a). Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B* **44**, 322–352.
- Shafer, G. (1982b). Lindley's paradox. *J. Amer. Statist. Assoc.* **77**, 325–351.
- Shafer, G. (1982c). Bayes's two arguments for the rule of conditioning. *Ann. Statist.* **10**, 1075–1089.
- Shafer, G. (1984). Conditional probability. Unpublished manuscript.
- Shafer, G. (1987). Probability judgment in artificial intelligence and expert systems. *Statist. Sci.* **2**, 3–16.
- Shannon, C.E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. U. Illinois Pr., Urbana.
- Shilov, G.E. and Gurevich, B.L. (1977). *Integral, Measure and Derivative: A Unified Approach*. Dover, New York.
- Shimony, A. (1955). Coherence and the axioms of confirmation. *J. Symbolic Logic* **20**, 1–28.
- Shore, J.E. and Johnson, R.W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy. *IEEE Trans. Inform. Theory*, **IT-26**, 26–37.
- Shortliffe, E.H. (1976). *Computer-Based Medical Consultations: MYCIN*. Elsevier, New York.
- Shortliffe, E.H. and Buchanan, B.G. (1975). A model of inexact reasoning in medicine. *Math. Biosciences* **23**, 351–379.
- Sierpinski, W. (1938). Fonctions additives non complètement additives et fonctions non mesurables. *Fund. Math.* **30**, 96–99.
- Simon, H.A. (1955). A behavioral model of rational choice. *Quarterly J. Econ.* **69**, 99–118.
- Simon, H.A. (1978). Rationality as process and as product of thought. *Amer. Economic Rev.* **68**, 1–16.
- Simon, H.A. (1982). *Models of Bounded Rationality*. (Two volumes.) MIT Pr., Cambridge (MA).
- Simon, H.A. (1983). *Reason in Human Affairs*. Stanford U. Pr.
- Skinner, B.F. (1938). *The Behavior of Organisms*. Appleton-Century-Crofts, New York.
- Skinner, B.F. (1945). The operational analysis of psychological terms. *Psychol. Rev.* **52**, 270–277.
- Skinner, B.F. (1953). *Science and Human Behavior*. Macmillan, New York.
- Skinner, B.F. (1971). *Beyond Freedom and Dignity*. Vintage, New York.
- Skyrms, B. (1980). Higher order degrees of belief. In *Prospects for Pragmatism*, ed. D.H. Mellor. Cambridge U. Pr.
- Slovic, P., Fischhoff, B. and Lichtenstein, S. (1977). Behavioral decision

- theory. *Ann. Rev. Psychol.* **28**, 1–39.
- Slovic, P. and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organiz. Behav. Hum. Perform.* **6**, 649–744.
- Slovic, P. and Lichtenstein, S. (1983). Preference reversals: a broader perspective. *Amer. Economic Rev.* **73**, 596–605.
- Smets, P., Mamdani, A., Dubois, D. and Prade, H. (eds.) (1988). *Non-standard Logics for Automated Reasoning*. Academic Pr., London.
- Smith, A.F.M. (1984). Bayesian statistics (with discussion). *J. Roy. Statist. Soc. Ser. A* **147**, 245–259.
- Smith, C.A.B. (1961). Consistency in statistical inference and decision (with discussion). *J. Roy. Statist. Soc. Ser. B* **23**, 1–37.
- Smith, C.A.B. (1965). Personal probability and statistical analysis (with discussion). *J. Roy. Statist. Soc. Ser. A* **128**, 469–499.
- Smith, L.D. (1986). *Behaviorism and Logical Positivism*. Stanford U. Pr.
- Solovay, R.M. (1970). A model of set theory in which every set of reals is Lebesgue measurable. *Ann. Math.* **92**, 1–56.
- de Sousa, R. (1971). How to give a piece of your mind; or the logic of belief and assent. *Rev. Metaphysics* **25**, 52–79.
- Spetzler, C.S. and Stael von Holstein, C.S. (1975). Probability encoding in decision analysis. *Management Sci.* **22**, 340–358.
- Spiegelhalter, D.J. (1986). A statistical view of uncertainty in expert systems. In *Artificial Intelligence and Statistics*, ed. W.A. Gale. Addison-Wesley, Reading (MA).
- Spiegelhalter, D.J. and Knill-Jones, R.P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology (with discussion). *J. Roy. Statist. Soc. Ser. A* **147**, 35–76.
- Stallings, W. (1977). Fuzzy set theory versus Bayesian statistics. *IEEE Trans. SMC-7*, 216–219.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp.*, Vol. 1, 197–206, eds. J. Neyman and E. L. Scott. U. California Pr., Berkeley.
- Stephanou, H.E. and Sage, A.P. (1987). Perspectives on imperfect information processing. *IEEE Trans. SMC-17*, 780–798.
- Stock, M. (1987). AI and expert systems: an overview. *AI Applications* **1**, 9–17.
- Stoer, J. and Witzgall, C. (1970). *Convexity and Optimization in Finite Dimensions*. Springer-Verlag, Berlin.
- Stone, M. (1963). The posterior *t* distribution. *Ann. Math. Statist.* **34**, 568–573.
- Stone, M. (1965). Right Haar measure for convergence in probability to quasi posterior distributions. *Ann. Math. Statist.* **36**, 440–453.

- Stone, M. (1976). Strong inconsistency from uniform priors (with discussion). *J. Amer. Statist. Assoc.* **71**, 114–125.
- Stone, M. (1982). Review and analysis of some inconsistencies related to improper priors and finite additivity. In *Logic, Methodology and Philosophy of Science 6*, eds. L.J. Cohen, J. Los, H. Pfeiffer and K-P. Podewski. North-Holland, Amsterdam.
- Stone, M. and Dawid, A.P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika* **59**, 369–373.
- Sudderth, W.D. (1980). Finitely additive priors, coherence and the marginalization paradox. *J. Roy. Statist. Soc. Ser. B* **42**, 339–341.
- Suppes, P. (1974). The measurement of belief (with discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 160–191.
- Suppes, P. (1975). Approximate probability and expectation of gambles. *Erkenntnis* **9**, 153–161.
- Suppes, P. (1984). *Probabilistic Metaphysics*. Blackwell, Oxford.
- Suppes, P. and Zanotti, M. (1977). On using random relations to generate upper and lower probabilities. *Synthese* **36**, 427–440.
- Szolovits, P. and Pauker, S.G. (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence* **11**, 115–144.
- Tanaka, H., Okuda, T. and Asai, K. (1976). A formulation of fuzzy decision problems and its application to an investment problem. *Kybernetes* **5**, 25–30.
- Taylor, C. (1964). *The Explanation of Behaviour*. Humanities Pr., New York.
- Taylor, S.J. (1973). *Introduction to Measure and Integration*. Cambridge U. Pr.
- Teller, P. (1973). Conditionalization and observation. *Synthese* **26**, 218–258.
- Teller, P. (1976). Conditionalization, observation, and change of preference. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. I, 205–259, eds. W.L. Harper and C.A. Hooker. Reidel, Dordrecht.
- Thole, U., Zimmerman, H.J. and Zysno, P. (1979). On the suitability of minimum and product operators for the intersection of fuzzy sets. *Fuzzy Sets and Systems* **2**, 167–180.
- Thorp, J., McClure, J. and Fine, T.L. (1982). The use of expert opinion in forecasting production cost of an electric utility. In *Proc. 1982 IEEE Internat. Large Scale Systems Symp.* IEEE Pr., New York.
- Tolman, E.C. (1949). *Purposive Behavior in Animals and Men*. U. Calif. Pr., Berkeley.
- Tribus, M. (1969). *Rational Descriptions, Decisions and Designs*. Pergamon Pr., New York.
- Tuomela, R. (ed.) (1978). *Dispositions*. Reidel, Dordrecht.
- Tversky, A. (1969). Intransitivity of preferences. *Psychol. Rev.* **76**, 31–48.

- Ulam, S. (1930). Zur Masstheorie in der allgemeinen Mengenlehre. *Fund. Math.* **16**, 141–150.
- Valentine, E.R. (1982). *Conceptual Issues in Psychology*. George Allen & Unwin, London.
- Vasely, W.E. and Rasmussen, D.M. (1984). Uncertainties in nuclear probabilistic risk analysis. *Risk Analysis* **4**, 313–322.
- Venn, J. (1888). *The Logic of Chance*, 3rd edition. Macmillan, London.
- Vickers, J. (1965). Coherence and the axioms of confirmation. *Philos. Sci.* **32**, 32–38.
- Vickers, J.M. (1976). *Belief and Probability*. Reidel, Dordrecht.
- Villegas, C. (1964). On qualitative probability σ -algebras. *Ann. Math. Statist.* **35**, 1787–1796.
- Villegas, C. (1971). On Haar priors. In *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott. Holt, Rinehart and Winston, Toronto.
- Villegas, C. (1977). Inner statistical inference. *J. Amer. Statist. Assoc.* **72**, 453–458.
- Villegas, C. (1981). Inner statistical inference II. *Ann. Statist.* **9**, 768–776.
- Wallace, D.L. (1959). Conditional confidence level properties. *Ann. Math. Statist.* **30**, 864–876.
- Walley, P. (1981). Coherent lower (and upper) probabilities. *Statistics Research Report* **22**. Univ. of Warwick, Coventry.
- Walley, P. (1982). The elicitation and aggregation of beliefs. *Statistics Research Report* **23**. Univ. of Warwick, Coventry.
- Walley, P. (1984–85). *Rationality and Vagueness*. Earlier draft of this book.
- Walley, P. (1987). Belief function representations of statistical evidence. *Ann. Statist.* **15**, 1439–1465.
- Walley, P. and Campello de Souza, F.M. (1986). The economic viability of solar energy systems. Technical report, UFPE and Cornell U. Revised version to appear in *Energy Systems and Policy*.
- Walley, P. and Fine, T.L. (1979). Varieties of modal (classificatory) and comparative probability. *Synthese* **41**, 321–374.
- Walley, P. and Fine, T.L. (1982). Towards a frequentist theory of upper and lower probability. *Ann. Statist.* **10**, 741–761.
- Walley, P. and Pericchi, L.R. (1988). Credible intervals: how credible are they? *Report no. 143*, Applied Math. Division, DSIR, Wellington (N.Z.). Revised version to appear in *Internat. Statist. Review*.
- Wallsten, T.S. and Budescu, D.V. (1983). Encoding subjective probabilities: a psychological and psychometric review. *Management Sci.* **29**, 151–174.
- Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R. and Forsyth, B. (1986). Measuring the vague meanings of probability terms. *J. Exper. Psychology: General* **115**, 348–365.

- Waterman, D.A. (1986). *A Guide to Expert Systems*. Addison-Wesley, Reading (MA).
- Watson, J.B. (1913). Psychology as the behaviorist sees it. *Psychol. Rev.* **20**, 158–177.
- Watson, J.B. (1930). *Behaviorism*. Norton, New York.
- Watson, S.R. (1974). On Bayesian inference with incompletely specified prior distributions. *Biometrika* **61**, 193–196.
- Watson, S.R., Weiss, J.J. and Donnell, M.L. (1979). Fuzzy decision analysis. *IEEE Trans. SMC-9*, 1–9.
- Weatherford, R. (1982). *Philosophical Foundations of Probability Theory*. Routledge and Kegan Paul, London.
- Weisweiller, R. (ed.) (1986). *Arbitrage*. Wiley, New York.
- White, A.R. (ed.) (1968). *The Philosophy of Action*. Oxford U. Pr.
- White, A.R. (1975). *Modal Thinking*. Cornell U. Pr., Ithaca (NY).
- Whittle, P. (1970). *Probability*. Penguin, Baltimore.
- Wilkinson, G.N. (1977). On resolving the controversy in statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 119–171.
- Williams, P.M. (1975a). Notes on conditional previsions. *Research Report*, School of Math. and Phys. Sci., Univ. of Sussex.
- Williams, P.M. (1975b). Coherence, strict coherence and zero probabilities. Contributed paper, *Fifth Internat. Congress of Logic, Methodology and Philos. Sci.* VI-29, 30.
- Williams, P.M. (1976). Indeterminate probabilities. In *Formal Methods in the Methodology of Empirical Sciences*, 229–246, eds. M. Przelecki, K. Szaniawski, and R. Wojcicki. Reidel, Dordrecht.
- Williams, P.M. (1978). On a new theory of epistemic probability. *Brit. J. Philos. Sci.* **29**, 375–387.
- Williams, P.M. (1980). Bayesian conditionalisation and the principle of minimum information. *Brit. J. Philos. Sci.* **31**, 131–144.
- Winkler, R.L. (1967a). The assessment of prior distributions in Bayesian analysis. *J. Amer. Statist. Assoc.* **62**, 776–800.
- Winkler, R.L. (1967b). The quantification of judgement: some methodological suggestions. *J. Amer. Statist. Assoc.* **62**, 1105–1120.
- Winkler, R.L. (1969). Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* **64**, 1073–1078.
- Winkler, R.L. (1971). Probabilistic prediction: some experimental results. *J. Amer. Statist. Assoc.* **66**, 675–685.
- von Winterfeldt, D. and Edwards, W. (1986). *Decision Analysis and Behavioral Research*. Cambridge U. Pr.
- Wolfenson, M. (1979). *Inference and Decision Making Based on Interval-Valued Probability*. Ph.D. Dissertation, Cornell U., Ithaca (NY).

- Wolfenson, M. and Fine, T.L. (1982). Bayes-like decision making with upper and lower probabilities. *J. Amer. Statist. Assoc.* **77**, 80–88.
- Wong, Y.-C. and Ng, K.-F. (1973). *Partially Ordered Topological Vector Spaces*. Clarendon Pr., Oxford.
- Yates, F. (1964). Fiducial probability, recognisable sub-sets and Behrens' test. *Biometrics* **20**, 343–360.
- Yates, J.F. and Zukowski, L.G. (1976). Characterization of ambiguity in decision making. *Behav. Sci.* **21**, 19–25.
- Zadeh, L.A. (1965). Fuzzy sets. *Inform. Control* **8**, 338–353.
- Zadeh, L.A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. SMC-3*, 28–44.
- Zadeh, L.A. (1975). Fuzzy logic and approximate reasoning. *Synthese* **30**, 407–428.
- Zadeh, L.A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* **1**, 3–28.
- Zadeh, L.A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems* **11**, 199–227.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.
- Zellner, A. (1977). Maximal data information prior distributions. In *New Developments in the Applications of Bayesian Methods*, 211–232, eds. A. Aykac and C. Brumat. North-Holland, Amsterdam.

Glossary of notation

This is a list of the notation that is used most frequently in the book. The list excludes some symbols that are used only ‘locally’, i.e., only in the section where they are first defined. In general, lower case Roman letters denote integers, statistical observations or density functions, upper case Roman denote single events, gambles or previsions, upper case Script denote sets (of events, gambles, linear previsions or possible observations), lower case Greek denote real numbers, possible states, statistical parameters or measures, and upper case Greek denote possibility spaces or functionals.

For special types of notation, see also sections 2.2.1 (concerning gambles), 2.7.1 (events), 3.3.1 (dominance), 6.2.1 and 6.2.5 (conditional previsions), and 7.3.1 (statistical models).

Roman letters

A, B, C, D	events (0–1 valued gambles)	81
$A(\cdot)$	multivalued mapping	182, 273
a	possible action	160
B	conditioning event	289
beta (s, t)	beta distribution with learning parameter s and mean t	218
$C, C(x)$	confidence interval estimator (in section 7.5)	378
D	draw (in a football game)	172
$\underline{E}, \bar{E}, E, \underline{E}(\cdot \mathcal{B})$	natural extension	112, 409
\exp	exponential function	
ext	set of extreme points	146, 613
$F, \bar{F}, \underline{F}$	distribution function, upper and lower distribution functions	130, 203
f_0, f	probability density function, joint density function	331

GLOSSARY OF NOTATION

675

$f(\cdot \theta)$	sampling density function	393
$G(X), G(X B), G(X \mathcal{B})$	marginal gambles	68, 289, 291
$g(\cdot x)$	posterior density function	369, 393
H	heads (in coin-tossing)	56
$H_P, \bar{H}, \underline{H}$	entropy, upper and lower entropy (section 5.12)	266, 541
h	prior density function	369
I	identity gamble	470
i, j, k, m, n	non-negative integers	
\inf	infimum	58
int	topological interior	
L	loss (in a football game)	172
L_x, L	likelihood function, or lower likelihood function	400, 431, 436–37
l	lower density function	200
\lim, \liminf, \limsup	limit, lower limit, upper limit	
\log	natural logarithm	
m	number of successes in Bernoulli trials	217
\max	probability assignment (section 5.13)	273
\min	maximum	
$N(\mu, \sigma^2)$	minimum	
n	Normal distribution with mean μ and variance σ^2	
P, Q	number of gambles in coherence	68, 73
P_0	condition	
P_T	sample size	217
$\underline{P}, \bar{Q}, \underline{R}$	linear prevision, additive	
\bar{P}	probability	66, 86, 89
\underline{P}_0	distinguished linear prevision	202, 229, 266
$\underline{P}_1, \bar{P}_n, \underline{P}^B, \bar{P}^x$	ideal linear prevision	105, 254, 356
$P(\cdot \cdot)$	lower prevision, lower probability	61, 82
$\bar{P}(\cdot \cdot), \bar{P}(\cdot \cdot)$	upper prevision, upper probability	64, 70, 82
$P(\cdot \Theta), \underline{P}(\cdot \theta)$	prior lower prevision	178
$\underline{P}(\cdot \mathcal{X}), \bar{P}(\cdot x)$	new lower prevision (after observing data, B or x)	178, 217, 285, 274
	linear conditional prevision	304
	conditional lower and upper previsions	289, 291
	statistical sampling model	349
	posterior lower previsions	362

$\bar{P}_B(\cdot \mathcal{B}), \underline{P}_B(\cdot \mathcal{B})$	conditional upper and lower prevision, modified on B	302, 390
$q(x)$	marginal density	331, 393
R_n	region of parameter values	218
$\underline{R}(\cdot B)$	regular extension (Appendix J)	639
r_n	relative frequency	221
S	scoring rule	247, 631
$S(Y)$	Θ -support of Y	363
$S_i(Y)$	\mathcal{B}_i -support of Y	344
s, t	parameters of beta distribution	218
s^2	sample variance	373
sup	supremum	58
T	tails (in coin-tossing)	56
$T(Y)$	\mathcal{X} -support of Y	363
U_x	upper likelihood function (section 8.5)	431
u	upper density function	200
$V_p, \bar{V}, \underline{V}$	variance under P , upper and lower variance (Appendix G)	617
W	win (in a football game)	172
X, Y, Z, U, V, W	gambles	58
X^*	evaluation functional	609
x, y	possible observations, especially statistical data	362
\bar{x}	vector of observations	373
\bar{x}	sample mean	
$x_\rho, \bar{x}_\rho, \underline{x}_\rho$	ρ -quantile, upper and lower ρ -quantile	204
Z^+, Z^-	positive and negative parts of Z	95
Script letters		
\mathcal{A}	a set of events, often a field or filter of events	82, 90, 100
\mathcal{B}	a partition of the possibility space	289
\mathcal{C}	a σ -field of subsets of Θ	363
\mathcal{D}	a set of almost-desirable gambles	151
\mathcal{D}^+	a set of strictly desirable gambles	155
\mathcal{E}	natural extension of \mathcal{D}	151
\mathcal{F}	domain of statistical model or joint prevision	363

\mathcal{G}	semigroup of transformations (section 3.5)	139
$\mathcal{G}(\mathcal{B})$	the set of all \mathcal{B} -measurable gambles	291
$\mathcal{H}, \mathcal{H}(B)$	domain of conditional prevision	289, 293
\mathcal{I}	a set of probable events (section 4.4)	189
\mathcal{I}^*	natural extension of \mathcal{I} (section 4.4)	189
\mathcal{J}	a set of structural judgements (section 9.7)	474
$\mathcal{K}, -\mathcal{K}$	domain of lower or upper prevision, especially prior prevision	61, 70, 388
$\mathcal{K}(\mathcal{A})$	the set of all \mathcal{A} -measurable gambles	129
$\mathcal{L}, \mathcal{L}(\Omega)$	the set of all gambles	58
$\mathcal{M}, \mathcal{M}(\underline{P})$	a set of linear previsions	132
\mathcal{P}	the set of all linear previsions on \mathcal{L}	132
\mathcal{Q}	the set of 0–1 valued additive probabilities	147
\mathcal{R}	a set of really desirable gambles (Appendix F)	614
\mathcal{S}	a σ -field of subsets of \mathcal{X}	363
\mathcal{T}	a linear space of gambles	608
\mathcal{V}, \mathcal{W}	convex sets	611
\mathcal{X}, \mathcal{Y}	statistical sample spaces	362
\mathcal{Z}	a linear space (Appendices D, E)	608
Greek letters		
$\alpha, \beta, \delta, \varepsilon, \zeta, \eta, \lambda, \mu, \rho, \sigma, \tau$	real numbers, constant gambles	58
Γ	an index set	
γ	{ index variable	
Δ, δ	{ confidence coefficient (section 7.5)	378
ζ	degree of imprecision	210, 329
η^2	stakes function (Appendices H, I)	629
Θ	prior variance	398
θ	parameter space	349
i	statistical parameter	349
κ	amount of information (section 5.3)	222
Λ	degree of conflict (section 5.4)	224
λ	linear functional	609
	stake	69, 630
	{ buying or selling price for gamble	61, 630
	location parameter, especially	

μ	$\left\{ \begin{array}{l} \text{Normal mean} \\ \text{fuzzy membership function} \\ \text{(section 5.11)} \end{array} \right.$	372–3, 398
ν	a measure, often Lebesgue measure	261
ξ	a measure	98, 331, 331
Π	bivariate functional (Appendices E, G)	613, 618
π	$\left\{ \begin{array}{l} \text{real number } 3.14159\dots \\ \text{permutation (Ch. 9)} \end{array} \right.$	457
Σ	summation	
σ	scale parameter, especially Normal standard deviation	372–73
$\bar{\sigma}, \underline{\sigma}$	upper and lower standard deviations (Appendix G)	617
τ	tax rate	96
Υ	variance functional (Appendix G)	618
v	possible state, bias of measuring instrument	399
Φ	$\left\{ \begin{array}{l} \text{Normal distribution function} \\ \text{possibility space (section 5.13)} \end{array} \right.$	275
ϕ	possible state, hyperparameter	260, 275
Ψ	possibility space	181, 184, 273
ψ	possible state, statistical parameter	273, 357, 468
Ω	possibility space	54
ω	possible state	54

Mathematical symbols

\mathbb{A}	set of possible actions	160
\mathbb{B}	Borel σ -field	98
\mathbb{R}	the set of real numbers	
\mathbb{R}^+	the set of positive reals	
\mathbb{R}^n	n -dimensional Euclidean space	
\mathbb{Z}^+	the set of positive integers	
$\{\gamma\cdot\}$	a set	
(α, β) or $[\alpha, \beta]$	open or closed interval of real numbers	
\times	product (of sets, σ -fields or linear previsions)	362, 453
\in	is an element of	
\notin	is not an element of	

\supset, \subset	set inclusion (not necessarily strict)	81
\cup	union	81
\cap	intersection	81
\complement	complementary set	28, 81
Δ	set-theoretic difference	81
\emptyset	empty set	81
\forall	universal quantifier ('for every')	
\vee	maximum	76–77
\wedge	minimum	76–77
$ \cdot $	absolute value or cardinality	
$\ \cdot\ $	supremum norm	609
$=, \neq$	equality, inequality	
\geq, \leq	dominance	58, 78
$>, <$	greater than, less than	
\geq^*, \leq^*	almost-preference, comparative probability	153, 191
\geq^*	natural extension of comparative probability (section 4.5)	192
$>$	strict preference	155
∞	infinity	
\propto	is proportional to	
\rightarrow	convergence or mapping	
\Leftrightarrow	equivalence	472
\int	integral	
$\binom{n}{x}$	binomial coefficient	374
\blacklozenge	end of proof	

Index of axioms

Unconditional lower previsions

- avoiding sure loss 68
- coherence 73
- coherence on linear spaces
 (axioms P1–P3) 63
- coherent upper previsions 65, 75
- monotonicity (P4) 76
- \emptyset -conglomerability (P7) 317
- full conglomerability (P8) 317
- regularity (P9) 640
- n*-coherence 599, 600

Lower probabilities

- coherence 84
- n*-coherence 600
- super-additivity 30
- 2-monotonicity 130
- complete monotonicity 272

Linear previsions

- general definition 86
- linear previsions on linear spaces
 (P0–P2, P5, P6) 66, 88–89
- additive probability 89–90
- additive probability on a field 90
- countable additivity 323

Axioms of precision

- linearity (P0) 66, 88
- self-conjugacy 87, 90
- finite additivity 90
- completeness (D8, D9, R9, R10)
 157, 195

Desirability

- axioms for almost-desirability
 (D0–D4) 60, 152
- strict desirability (D2, D3, D5–D7)
 155
- real desirability (D2, D3, D10–D12)
 614
- completeness (D8, D9) 157

Preference

- axioms for almost-preference
 (R0–R5) 154
- strict preference (R2, R3, R5–R8)
 156
- real preference (R1–R3, R5, R11,
 R12) 616
- transitivity (R3) 29, 154
- completeness (R9, R10) 157
- n*-coherence and de Finetti's
 axioms for comparative
 probability 600–1

INDEX OF AXIOMS

Conditional previsions

- separate coherence (C1–C3) 292
- avoiding sure loss 294
- coherence 294
- coherence on linear spaces (C4–C6)
 302
- coherence when \mathcal{H} contains
 \mathcal{H} (C7–C9) 302
- coherence when \mathcal{H} contains
 \mathcal{H} (C10–C12) 303
- generalized Bayes rule (C12)
 297, 303
- Bayes' rule 298, 305
- coherence for linear previsions
 (C13–C15) 305
- conglomerative axioms (C5, C8,
 C11, C13–C15) 302–03, 305
- regularity (C16) 640

General concepts of coherence

- avoiding uniform sure loss 343
- avoiding partial loss 344
- avoiding sure loss 346
- avoiding uniform loss 346
- weak coherence 346
- general definition of coherence
 346–47
- W-coherence 642

Statistical models

- coherence of sampling model and
 posterior previsions (S1–S5)
 363–64, 366
- coherence of sampling model,
 prior and posterior previsions
 (S6, S7) 389, 391
- coherence of sampling model,
 joint prior and posterior
 previsions (S5, C11, C12) 404
- coherence of Bayesian models
 (S1, S6*) 394
- generalized Bayes rule (statistical
 versions) 400, 427, 432
- coherence of inferences from
 generalized Bayes rule (S5, C11)
 401
- \mathcal{X} -conglomerability 317
- coherence of confidence interval
 estimators (I1–I4) 380

Author index

Aczel, J. 523, 534, 540
 Adams, J.B. 50
 Agassi, J. 500
 Allais, M. 492
 Amaral-Turkman, M.A. 523
 Anderson, B.D.O. 536
 Anderson, J.R. 482
 Andrews, S.E. 490
 Anger, B. 502
 Anscombe, F.J. 241, 533
 Aristotle 484, 485
 Armstrong, D.M. 481, 482, 491
 Armstrong, T.E. 516
 Asai, K. 538
 Aumann, R.J. 241, 490, 533
 Ayer, A.J. 487, 501
 Barnard, G.A. 374, 501, 558, 559, 568, 584, 585
 Barndorff-Nielsen, O. 570
 Barnett, V. 489, 568
 Baron, J. 484
 Basu, D. 583, 584
 Bayes, T. 43, 220, 228, 377, 493, 521
 Beach, B.H. 501
 Becker, S.W. 48
 Bellman, R.E. 538, 539
 Beran, R.J. 489, 545
 Berberian, S.K. 504, 608
 Berger, J.O. 43, 44, 255, 256, 257, 374, 478, 479, 487, 488, 492, 497, 500, 501, 507, 508, 510, 517, 518, 520, 525, 526, 527, 528, 529, 530, 532, 535, 536, 538, 540, 569, 570, 573, 583, 584
 Berliner, L.M. 497, 518, 536
 Bernardo, J.M. 228, 229, 233, 525, 526, 528, 540, 583, 584
 Bernoulli, J. 489, 543

Beyth-Marom, R. 514, 539
 Bhaskara Rao, K.P.S. 496, 498, 504
 Bhaskara Rao, M. 496, 498, 504
 Biller, W.F. 538
 Billingsley, P. 498, 553, 558, 559, 574, 587, 588
 Birnbaum, A. 483, 572, 583, 584
 Blachman, N.M. 523
 Black, M. 514, 539
 Blackburn, S. 563
 Blackwell, D. 528
 Block, N. 482, 491
 Blum, J.R. 530, 536
 Boes, D.C. 569
 Boesky, I.F. 495
 Bondar, J.V. 569
 Boole, G. 43, 44, 487, 489, 494, 503, 526, 586
 Borel, E. 43, 46, 481, 489, 624, 625
 Box, G.E.P. 43, 228, 229, 233, 500, 525, 526, 528, 538, 564, 565, 567, 568, 583, 584, 593
 Braithwaite, R.B. 481, 501
 Breiman, L. 553
 Bridgman, P.W. 482
 Brown, L.D. 383, 573
 Brown, R.V. 499, 500, 534, 536, 537
 Brownson, F.O. 48
 Buchanan, B.G. 50, 490, 536
 Budescu, D.V. 48, 510, 539, 672
 Buehler, R.J. 47, 377, 381, 383, 494, 506, 530, 568, 570
 Burgess, J.P. 514
 Campello de Souza, F.M. viii, 47, 488, 500, 508, 520, 529
 Cano, J.A. 517, 518, 536
 Caramazza, A. 539
 Carnap, R. 16, 23, 34, 43, 113, 238, 480,

AUTHOR INDEX

483, 485, 486, 487, 494, 501, 514, 515, 521, 524, 525, 547, 565
 Casella, G. 569, 571, 572
 Chamberlain, G. 536
 Chernoff, H. 491
 Choquet, G. 502, 505
 Chuang, D.T. 536
 Cohen, L.J. 39, 50, 488, 490, 501, 539
 Cornfield, J. 569, 571
 Cox, D.R. 386, 483, 487, 489, 500, 525, 526, 527, 528, 565, 567, 569, 570, 572, 573, 583, 591
 Cox, R.T. 44, 534
 Curley, S.P. 48
 Dalal, S.R. 518
 Daniell, P.J. 493
 Daroczy, Z. 523, 540
 Darwall, S.L. 484, 485, 500
 Davis, M. 507
 Dawid, A.P. 361, 488, 500, 527, 565, 566, 568, 587, 590, 591, 592
 Day, J.P. 514
 Day, M.M. 504, 505, 608
 De Morgan, A. 43, 479
 DeGroot, M.H. 43, 45, 241, 516, 518, 523, 528, 532, 647
 Dempster, A.P. 11, 44, 47, 182, 208, 272, 273, 274, 275, 281, 358, 479, 489, 511, 513, 536, 542, 543, 544, 545, 564, 568
 Dennett, D.C. 18, 481, 482, 491, 529
 DeRobertis, L. 498, 517, 535, 536, 576
 Dewey, J. 484
 Diaconis, P. 518, 544, 547, 559, 560
 Dias, P. 540
 Dickey, J.M. 44, 499, 528, 536, 537, 568
 Donnell, M.L. 50, 262, 538, 672
 Dretske, F. 482
 Dubins, L.E. 313, 496, 528, 552, 554, 557
 Dubois, D. 538, 669
 Duda, R.O. 49
 Dunford, N. 608
 Dunsmore, I.R. 523
 Edwards, A.W.F. 524, 568
 Edwards, W. 43, 44, 480, 488, 492, 501, 510, 528, 533, 536, 538, 546, 623
 Eells, E. 485
 Einhorn, H.J. 44, 48, 480, 532, 535, 536
 Ellis, R.L. 43, 480, 526
 Ellsberg, D. 48, 530, 532
 Elster, J. 484, 485, 532
 Epstein, R.A. 602
 Farquhar, P.H. 492
 Feagans, T.B. 538
 Feddersen, A.P. 383
 Feller, W. 526, 541, 591
 Fellner, W. 44, 47, 48, 520, 521, 535, 536
 Ferguson, T.S. 508, 509
 Fine, T.L. ix, 44, 45, 46, 47, 49, 478, 479, 480, 487, 489, 491, 496, 501, 507, 514, 515, 516, 517, 525, 526, 528, 531, 532, 533, 539, 540, 541, 542, 543, 547, 562, 563, 564, 565, 586, 587, 588, 589, 591, 592, 600
 de Finetti, B. viii, ix, 3, 16, 20, 43, 44, 45, 54, 55, 57, 66, 81, 86, 111, 138, 241, 242, 243, 246, 247, 282, 295, 320, 321, 322, 326, 327, 361, 365, 443, 460, 463, 465, 466, 478, 479, 480, 481, 482, 483, 484, 486, 487, 488, 489, 490, 491, 493, 494, 496, 497, 498, 499, 500, 501, 503, 507, 510, 511, 512, 513, 515, 516, 528, 531, 532, 533, 534, 537, 539, 541, 545, 546, 548, 549, 552, 553, 554, 555, 556, 557, 558, 559, 560, 566, 567, 580, 585, 586, 587, 589, 590, 591, 592, 601, 624, 632, 644
 Fischhoff, B. 480, 510, 662
 Fishburn, P.C. 46, 241, 489, 490, 499, 507, 515, 521, 531, 532, 533, 536
 Fisher, R.A. 44, 232, 362, 373, 382, 483, 488, 521, 522, 524, 526, 528, 536, 562, 568, 569, 570, 571
 Fodor, J.A. 482, 491
 Folks, J.L. 536
 Folks, L. 386, 483, 558, 559, 562, 569, 572, 573, 584
 Forsyth, B. 672
 Forsyth, R. 49, 538
 Fraser, D.A.S. 373, 501, 568, 570, 571, 585
 Freedman, D.A. 579
 Freeling, A.N.S. 50, 263, 538, 539, 540
 French, S. 516

- Gaines, B.R. 538, 539
 Gale, D. 511, 612
 Gale, W.A. 490
 Gardenfors, P. 44, 49, 507, 530, 535, 538
 Gardner, H. 482
 Gauss, C.F. 527, 567, 569
 Geisser, S. 521, 526
 Gerber, H.U. 498
 Giere, R.N. 480, 562
 Giertz, M. 539
 Giles, R. 47, 499, 500, 503, 506, 510, 519, 539
 Gillies, D.A. 480, 562
 Giron, F.J. 47, 506, 507, 528, 536
 Goel, P.K. 523
 Goldstein, M. 536, 546, 547, 549, 552, 555, 557, 560, 592
 Good, I.J. ix, 43, 44, 46, 255, 256, 257, 478, 479, 481, 484, 485, 487, 488, 500, 501, 509, 519, 523, 524, 525, 528, 529, 534, 535, 536, 537, 538, 540, 551, 579, 589, 590, 591, 592
 Granirer, E. 504
 Graybill, F.A. 569
 Greenleaf, F.P. 504
 Griffiths, A.P. 481
 Grize, Y.-L. 49, 592
 Gurevich, B.L. 493
 Hacking, I. 478, 479, 480, 481, 482, 489, 491, 493, 526, 543, 545, 547, 559, 562, 563, 568, 569, 592
 Hagen, O. 492
 Hailperin, T. 489, 503
 Haldane, J.B.S. 223, 228, 526, 527, 542
 Hall, W.J. 228, 518, 525
 Halmos, P.R. 502, 506, 588
 Hamblin, C.L. 514
 Hampton, J.M. 510
 Harsanyi, J.C. 519
 Hartigan, J.A. 230, 488, 498, 517, 527, 528, 535, 536, 568, 576
 Heath, D.C. 374, 390, 495, 496, 502, 557, 566, 567, 568, 571, 573, 575, 579, 590, 591, 602
 Hempel, C.G. 482
 Hernández, A. 517, 518, 536
 Hersh, H.M. 539
 Hewitt, E. 590, 591
 Hey, J.D. 519
 Hill, B.M. 485, 527, 528, 555, 556, 557, 558, 568
 Hinkley, D.V. 386, 483, 489, 500, 525, 526, 527, 528, 567, 569, 570, 572, 573, 583, 592
 Hisdal, E. 539
 Hodges, J.L. 501
 Hogarth, R.M. 44, 48, 480, 482, 488, 501, 510, 532, 535, 536, 623
 Holmes, R.B. 505, 608, 610, 611, 612, 613
 Hora, R.B. 568
 Huber, G.P. 510
 Huber, P.J. 44, 49, 479, 493, 494, 495, 496, 497, 502, 503, 505, 517, 518, 536, 563, 564
 Hull, C.L. 481
 Hunter, J.S. 565
 Hunter, W.G. 565
 Isaacs, H.H. 536
 Jain, R. 538
 Jameson, G.J.O. 506, 608
 Janis, I.L. 484
 Jaynes, E.T. 11, 23, 43, 228, 266, 267, 271, 486, 525, 526, 527, 528, 540, 541, 542, 590, 591, 592
 Jech, T.J. 506
 Jeffrey, R.C. 43, 339, 481, 483, 485, 490, 491, 492, 525, 533, 543, 544, 560
 Jeffreys, H. 16, 23, 43, 45, 113, 223, 228, 229, 232, 238, 241, 244, 245, 362, 373, 478, 486, 500, 520, 525, 526, 527, 528, 532, 567, 571
 Jenkins, G.M. 558, 646
 Jepson, C. 664
 Johnson, R.W. 540
 Johnson, W.E. 589
 Kadane, J.B. 323, 496, 536, 555, 557, 558, 559, 667
 Kahneman, D. 488, 501, 536
 Kanal, L.N. 490
 Kaplan, M. 515, 516, 588
 Karmarkar, N. 511
 Kekes, J. 484
 Kelley, J.L. 506, 608, 609, 610

- Kemeny, J. 494
 Kempthorne, O. 386, 483, 558, 559, 562, 569, 572, 573, 584, 585
 Kent, S. 514, 539
 Keynes, J.M. ix, 23, 34, 44, 45, 113, 238, 244, 486, 487, 489, 497, 515, 516, 522, 523, 526, 532, 585
 Kickert, W.J.M. 538, 539, 540
 Kiefer, J. 387, 570, 572
 Kingman, J.F.C. 590
 Kmietowicz, Z.W. 48, 507, 536
 Kneale, W. 514
 Knight, F.H. 209, 489, 519
 Knill-Jones, R.P. 50, 490
 Kolmogorov, A.N. 32, 282, 306, 307, 327, 489, 549, 553, 558, 561
 Koopman, B.O. 45, 515, 557
 Kraft, C. 515, 601
 Krantz, D.H. 515, 664
 Kudō, H. 517, 518, 536
 Kuhn, T.S. 478
 Kullback, S. 523
 Kumar, A. 49, 592
 Kunda, Z. 664
 Kunreuther, H. 48, 501, 532
 Kyburg, H.E. 34, 44, 48, 238, 479, 480, 484, 486, 487, 488, 501, 529, 531, 536, 562, 563, 590, 591, 592
 Lad, F.R. 482, 488, 499
 Lane, D.A. 555, 556, 566, 568, 573, 579
 Laplace, P.S. de 43, 220, 228, 377, 521, 522, 525, 527, 569
 Lauritzen, S.L. 49, 490
 Leamer, E.E. 44, 47, 492, 493, 518, 521, 525, 536, 565, 623
 Lehman, R.S. 494
 Lehmann, E.L. 501, 569, 570, 571, 573
 Leibler, R.A. 523
 Lemmer, J.F. 490
 Levi, I. 34, 44, 48, 238, 478, 480, 481, 482, 483, 487, 488, 490, 496, 499, 507, 519, 520, 530, 532, 536, 537, 543, 545, 547, 555, 557, 562, 563, 564, 592
 Levine, R.D. 540
 Lewis, D. 563
 Lichtenstein, S. 480, 488, 501, 510, 536
 Lindley, D.V. 16, 43, 44, 45, 241, 242, 245, 246, 478, 486, 487, 490, 492, 499, 500, 516, 521, 523, 526, 528, 532, 533, 534, 535, 536, 537, 538, 544, 546, 568, 583, 590, 647
 Lindman, H. 44, 528, 533, 536
 Loomis, L.H. 493
 Luce, R.D. 660
 Machlup, F. 524
 Mackenzie, B.D. 481
 Mallows, C.L. 488, 565, 590
 Mamdani, A. 669
 Mann, L. 484
 Mansfield, U. 524
 Manski, C.F. 491, 513, 517, 518, 536
 March, J.G. 488
 Martin, A.W. 538
 McClure, J. 47, 517
 Mellor, D.H. 44, 352, 354, 480, 481, 482, 483, 499, 500, 510, 521, 534, 536, 537, 545, 547, 559, 562, 563, 564, 624
 Menges, G. 530
 Mill, J.S. 43, 526
 von Mises, R. 44, 351, 480, 528, 536, 564
 Mitchell, A.F.S. 528
 Mood, A.M. 569
 Moore, G.H. 504, 505, 506
 Moore, P.G. 510
 Moreno, E. 517, 518, 536
 Morgenstern, O. 490, 492
 Moses, L.E. 491
 Mosteller, F. 487, 500, 536, 544, 565
 Murphy, A.H. 536
 Nachbin, L. 568
 Nagel, E. 565
 Namioka, I. 506, 608
 Nathanson, S. 484
 Nau, R.F. 489, 536, 538
 Nazaret, W.A. 538
 von Neumann, J. 490, 492, 613
 Neyman, J. 11, 22, 42, 112, 235, 377, 386, 483, 569, 572, 573
 Ng, K.-F. 506
 Nisbett, R.E. 483, 488
 Norman, R. 481
 Novick, M.R. 228, 521, 525, 590

- Okuda, T. 538
 Olshen, R.A. 571
 Papamarcou, A. 49, 564, 592
 Parikh, R. 507, 559
 Parnes, M. 507, 559
 Pauker, S.G. 490
 Pearl, J. 49, 489, 490, 538, 560
 Pearman, A.D. 48, 507, 536
 Pears, D.F. 484
 Pearson, K. 521
 Pedersen, J.G. 568, 569, 571
 Peirce, C.S. 43, 481, 500, 522, 523, 526, 562
 Pereira, C.A.B. 529
 Pericchi, L.R. ix, 479, 497, 500, 510, 517, 518, 525, 528, 529, 535, 536, 537, 538, 559, 568, 576, 584
 Perks, W. 228
 Phillips, L.D. 521, 583, 662
 Pierce, D.A. 377, 536, 570, 572, 573
 Pincus, D. 506
 Pitz, G.F. 480, 536
 Polasek, W. 518, 536
 Popper, K.R. 480, 500, 501, 522, 562, 563, 564, 565
 Potter, J.M. 536
 Prade, H. 50, 538, 539, 669
 Pratt, J.W. 515, 516, 533, 545, 547, 572, 601
 Purves, R.A. 579
 Putnam, H. 488
 Quinlan, J.R. 50, 513
 Raiffa, H. 241, 516, 518, 519, 521, 523, 531, 532, 533, 545, 547
 Ramsey, F.P. 43, 45, 241, 243, 244, 249, 481, 483, 488, 489, 490, 491, 494, 531, 534, 545, 548, 562
 Rapoport, A. 672
 Rasmuson, D.M. 532
 Rawls, J. 484, 485
 Raz, J. 481
 Regazzini, E. 549, 555, 642
 Reichenbach, H. 480, 534, 562
 Rényi, A. 549, 553
 Rescher, N. 514, 515
 Rios, S. 47, 506, 507, 528, 536
 Robinson, G.K. 377, 569, 570, 571, 572, 573
 Rockafellar, R.T. 511
 Rorty, R. 500
 Rosenblatt, J. 530, 536
 Rosenkrantz, R.D. 43, 267, 525, 528, 540, 541
 Ross, L. 483, 488
 Royden, H.L. 493, 498, 504, 559, 574, 588
 Rubin, H. 484
 Russell, B. 44, 499, 532, 543, 562
 Ryle, G. 19, 481, 565
 Sachs, N.J. 480, 536
 Sage, A.P. 490
 Sahlin, N. 44, 49, 507, 530, 535, 538
 Salmon, W.C. 480, 562
 Savage, L.J. 16, 43, 44, 45, 241, 242, 246, 247, 478, 483, 484, 485, 489, 490, 491, 492, 496, 500, 515, 516, 521, 528, 529, 531, 533, 534, 536, 537, 546, 547, 548, 557, 590, 591, 623, 631
 Scheife, P. 539
 Schervish, M.J. 323, 555, 556, 659
 Schlaifer, R. 516, 518, 521, 523, 533, 545, 547
 Schoemaker, P.J.H. 480, 536
 Schum, D.A. 490, 538
 Schwartz, J.J. 608
 Scott, D. 516
 Seidenberg, A. 515, 601
 Seidenfeld, T. 323, 555, 568, 569, 659, 667
 Sen, A.K. 529, 530
 Shackle, G.L.S. 50, 539
 Shafer, G. ix, 11, 23, 47, 208, 272, 273, 274, 275, 280, 281, 358, 483, 489, 490, 491, 513, 528, 542, 543, 544, 545, 546, 547, 548, 550, 555, 559, 560, 564
 Shannon, C.E. 268, 523, 540
 Shilov, G.E. 493
 Shimony, A. 494, 540
 Shore, J.E. 540
 Shortliffe, E.H. 49, 50, 490, 536
 Sierpinski, W. 506
 Simon, H.A. 484, 488, 530
 Skinner, B.F. 19, 482
 Skyrms, B. 537, 560
 Slovic, P. 480, 488, 501, 510, 536, 659

- Smets, P. 490
 Smith, A.F.M. 538, 591, 647
 Smith, C.A.B. ix, 44, 46, 106, 255, 482, 483, 485, 489, 490, 492, 494, 495, 501, 503, 506, 507, 531, 563
 Smith, L.D. 481
 Smokler, H.E. 488
 Snell, E.J. 488, 565
 Solovay, R.M. 505
 de Sousa, R. 510, 537
 Spetzler, C.S. 510, 516
 Spiegelhalter, D.J. 49, 50, 490
 Sprott, D.A. 568
 Stael von Holstein, C.S. 510, 516
 Stallings, W. 538
 Stein, C. 527
 Stephanou, H.E. 490
 Stock, M. 49
 Stoer, J. 608, 613
 Stone, M. 527, 528, 556, 566, 567, 568
 Strassen, V. 49, 497, 502, 563
 Suderth, W.D. 374, 390, 495, 496, 502, 527, 557, 566, 567, 568, 571, 573, 575, 579, 590, 591, 602
 Suppes, P. 44, 46, 480, 485, 489, 493, 499, 514, 515, 516, 521, 529, 533, 534, 563, 564, 600, 660
 Szolovits, P. 490
 Tanaka, H. 538
 Taylor, C. 481
 Taylor, S.J. 559, 588
 Teller, P. 545, 550
 Thole, U. 539
 Thomas, H. 510
 Thorp, J. 47
 Tiao, G.C. 43, 228, 229, 233, 500, 525, 526, 528, 538, 564, 567, 568, 583, 584
 Tolman, E.C. 481
 Tribus, M. 267, 487, 534, 540, 542, 544
 Tuomela, R. 481, 562, 563
 Tversky, A. 485, 488, 499, 500, 501, 534, 536, 537, 659, 660
 Ulam, S. 498, 557
 Valentine, E.R. 483
 Vasely, W.E. 532
 Venn, J. 43, 479, 480, 526
 Vickers, J.M. 485, 491
 Villegas, C. 228, 515, 516, 525, 532, 568
 Wallace, D.L. 487, 500, 536, 565, 569
 Walley, P. vii, viii, 44, 46, 47, 49, 478, 479, 488, 489, 495, 496, 497, 500, 502, 503, 508, 510, 512, 514, 515, 516, 517, 518, 519, 520, 528, 529, 535, 536, 537, 539, 542, 543, 545, 551, 559, 563, 564, 565, 568, 576, 582, 587, 589, 590, 591, 592
 Wallsten, T.S. 510, 539, 648
 Waterman, D.A. 49
 Watson, J.B. 19, 482
 Watson, S.R. 50, 262, 263, 264, 536, 538, 539
 Weatherford, R. 479
 Weaver, W. 523, 540
 Weinberg, S. 648
 Weiss, J.J. 50, 262, 538, 672
 Weisweiller, R. 495
 White, A.R. 481, 490, 514
 Whittle, P. 493, 497
 Wilkinson, G.N. 568
 Williams, P.M. ix, 13, 46, 106, 414, 489, 491, 493, 495, 499, 502, 503, 506, 519, 540, 542, 543, 544, 548, 549, 558, 561, 577, 614, 642, 643
 Winkler, R.L. 501, 510, 511, 521, 533, 536, 623, 632, 638
 Winsten, C.B. 558, 646
 von Winterfeldt, D. 43, 480, 488, 492, 501, 510, 536, 538, 546, 623
 Witzgall, C. 608, 613
 Wolfson, M. 47, 507, 517, 600
 Wolpert, R.L. 500, 536, 569, 570, 573, 583, 584
 Wong, Y.-C. 506
 Yates, F. 569, 571
 Yates, J.F. 48
 Ylvisaker, D. 518
 Zabell, S. 547, 559, 560
 Zadeh, L.A. 50, 261, 265, 519, 538, 539
 Zanotti, M. 489
 Zellner, A. 228, 229, 525, 528, 538, 584
 Zidek, J.V. 527, 568
 Zimmerman, H.J. 539
 Zukowski, L.G. 48
 Zwick, R. 672
 Zysno, P. 539

Subject index

Action 24, 160–61, 235–41
 admissible 161
 Bayes 162, 165–66, 237–40
 identified with gamble 160–61
 maximal 161–66
 minimax 163, 166, 240, 269–70, 619, 630
 optimal, *see* Action, Bayes
 reasonable but not Bayes 165–66
 satisfactory 239–41
see also Decision; Behaviour
 Additivity 89–92
 countable 323–27, 453–54
 de Finetti's arguments against 326
 as a rationality axiom 32, 326–27
 relation to conglomerability 310–13, 323–24
 finite versus countable, examples 70, 97–99, 310–11, 313, 321–23, 325–26, 365, 390, 396, 420–21, 450
 sub- and super- 30, 84, 600
 Admissibility 161, 231
 compared to coherence 345, 509, 527
 relation to Bayesian optimality 163–64, 253
 Aggregation, *see* Combination of assessments
 Agreement, interpersonal 111–13, 361
 Algorithms
 for checking coherence 597–98
 for computing extreme points 174–76, 511
 for computing joint previsions 316–17, 454–55
 for computing natural extension 136, 174–76
 for generalized Bayes rule 550, 581

Ambiguity 216, 261–64, 266
 Amenability 504
 Approximation in model building 117, *see also* Idealization
 Arbitrage 72
 Artificial intelligence, *see* Expert systems
 Assessment 32–42, 167
 of class of conjugate priors 205–06, 218–19, 221, 225
 combination of, *see* Combination of assessments
 in football experiment 634–35
 guided by an expert system 40, 512
 incomplete 173, 212, 215–16, 256–57, 285–86
 inconsistent 187–88, 213–14, 216–17
 of intervals of measures 201–02
 of neighbourhoods 202–03
 of prior previsions 198, 200–06, 221, 403, 424, 429
 reliable 285, 288–89
 of sampling models 359–60
 as a sequential process 177–78
 of upper and lower density functions 200–02
 of upper and lower distribution functions 203–05
 of upper and lower probabilities 197–99
 of upper and lower probability ratios 199–200
 of upper and lower quantiles 204–05
 using conditional previsions 286, 300
 versus elicitation 15, 22–23, 167
 well grounded in evidence 113–14
 see also Strategies, assessment; Judgement; Elicitation; Extension

SUBJECT INDEX

Assessment strategies, *see* Strategies, assessment
 Assignment, probability 273
 Avoiding sure loss 28–29, 67–72, 115–16, 293–94, 346
 in betting on horses 603–05
 comparison with coherence 67, 74
 in football experiment 635
 justification for 68, 295, 343–45
 objections to the emphasis on 114–16
 relation to dominating linear previsions 134–35
 see also Loss
 Axiom of choice 505, 611
 Axioms, index of 680
 Axioms, role of 31–32
 Banach limits 143–44
 Bayes decision rule 164
 Bayesian inference, *see* Inference, statistical, Bayesian
 Bayesian sensitivity analysis, *see* Sensitivity analysis
 Bayes' rule 185, 297–98, 305–06, 427–28
 for density functions 331–33, 393–96, 438–40
 in statistical problems 393–97, 427–9, 432
 not sufficient for coherence 305, 321
 as an updating strategy 336, 560
 not valid for arbitrary conditioning variables 332, 559
 Bayes rule, generalized (GBR) 185, 297–301, 400–03, 424–34
 algorithms for solving 550, 581
 as coherence relation (axiom C12) 297, 300, 303–04, 335
 conditions for applying 300–01, 334–37
 examples of 298–300, 308–13
 for imprecise likelihood functions 337–38, 430–34
 as a lower envelope 298, 429, 432
 as natural extension 318–19, 411–12, 424–27
 for precise likelihood functions 392, 400–03, 424–29
 as an updating strategy 285–86, 335–40
 see also Elicitation; Indeterminacy; Uncertainty
 Bernoulli models 205, 217, 467
 approximate 356, 454–55
 compared to exchangeability 112–13, 463–66
 robust 467–71
 Beta distribution, *see* Distribution, beta
 Betting
 on confidence intervals 380–82, 386
 on football games 632–38
 on horses 94–95, 245, 602–07
 objections to the emphasis on 114–16
 on parameter values 118–19, 367–68, 429
 pari-mutuel system 94–95, 130–31, 135, 147, 202–03, 604–06

Betting procedure, biased 381–82, 384, 565–66
 Betting rates 82, 602–03
 of bookmakers 87–88, 95, 245, 602, 605, 625–28
 fair (two-sided) 89, 243–44, 251, 624–25
 lower 82, 251, 632
 upper 82, 602, 632
 see also Price for gamble; Probability, additive; Probability, lower; Probability, upper
 Bias
 in elicitation 624–28
 of measuring instrument, model for 399–400, 405, 423–24
 Bookmakers (*see also* Betting rates) 95, 245, 625–28
 Borel σ -field 98
 Borel's paradox 328–30, 332, 450–51
 Bounds, upper and lower
 for chances 356–57, 468
 for conditional probabilities 296, 301
 for density functions 199–201
 for posterior previsions 432
 for probabilities 35, 46–47, 216
 see also Sensitivity analysis; Envelopes

Calibration 39
 Cantelli-Lévy paradox 365, 450
 Certainty 54, 209
 Certainty factors (MYCIN) 50
 Chance 351–59, 465–66
 imprecise 358–59, 454, 468
 lower bound for 356–57, 468, 588
 Choice 160–66, 236–41
 arbitrary 19, 23–41, 210
 between models for uncertainty 159–60
 strategies for 239–40
 versus preference 236–37, 242, 247–48
 see also Decision; Action; Preference
 Classification of probable events, *see* Probability, classificatory
 Coarsening (of a possibility space) 181–82
 Coherence 29–30, 63–66, 72–76, 346–47
 algorithm for checking 597–98

axioms, *refer to* Index of axioms
 680
 of Bayesian models 394, 405
 compared to avoiding sure loss 67, 74, 347
 of conditional with unconditional previsions 293–96, 301–05
 countable 32, 495, 556
 in football experiment 636
 general concepts 342–49
 justification for 60–61, 64, 73–74, 295–96, 347
 of lower previsions 72–76, 134
 of lower probabilities 84–86, 135, 600
 n- 599–601
 relation to dominating linear previsions 134–35, 145–46
 relation to natural extension 74, 122–23, 408–09
 separate 289–93
 for statistical models 366, 380, 391, 392, 397, 401, 404
 strength of 74, 116–18
 strict 32, 495, 581
 W- 577–78, 642–44
 weak 346–47, 391, 566
 for win-and-place betting on horses 605–06
 Coin tossing, models for 119, 274–75, 298–99, *see also* Bernoulli models
 Collective (von Mises) 351, 564
 Combination of assessments 30, 39, 170–72, 178–80, 186–88, 275–81, 338, 452–57
 Compactness 610
 of $M(P)$ 145–46, 505, 610
 Comparability, *see* Completeness
 Completeness 157, 190, 195–97, 241–48, 601
 Computations, *see* Algorithms
 Conclusions 21–22
 indeterminate 2, 5–6, 209–10, 226
 robust 5–6, 256
 see also Reasoning; Inference, statistical; Revision, posterior
 Conditioning 185–86, 282–340
 on continuous variables 330–33, 436–40

on events of probability zero 306–07, 328–34, 391–92, 615–16
 examples of 185–86, 298–300, 308–13, 321
 de Finetti's theory 282, 326–27
 in frequentist inference 382, 387–88
 Kolmogorov's theory 282, 306–08, 327
 see also Revision, conditional; Conglomerability; Bayes rule, generalized; Dempster's rule of conditioning; Jeffrey's rule; Updating
 Cone, convex 76
 Confidence coefficient 378
 data-dependent 387
 as a posterior lower probability 386
 as a posterior probability 22, 378–80, 388
 see also Confidence intervals
 Confidence intervals 377–88
 axioms for coherence 380
 examples of incoherence 379, 382–85
 interpretations of 378, 382, 385–87
 need for conditioning 382, 387–88
 posterior confidence in 378–80, 388
 zero confidence in 379, 385, 569
 Conflict 187–88, 213–14, 222–26
 between expert opinions 47, 213–14, 520
 between prior and statistical information 6, 213–14, 222–26
 between probability models 187–88, 213–14, 216–17
 degree of 223, 224–26
 Conglomerability 317–27, 614–16
 B- 317
 examples of failure 311, 321–22
 full 317, 614
 justification for 319–20
 relation to countable additivity 310–13, 323–24
 relation to sensitivity analysis 327
 for uncountable spaces 324–26
 Z- 424–27, 430–31
 Conglomerative principle, *see* Principles of rationality, conglomerative
 Conjugacy 64–65, 70, *see also* Self-conjugacy
 Conjugate prior, see Density function, natural-conjugate
 Conjunction rule 186
 Consensus, *see* Agreement, interpersonal; Combination of assessments
 Consistency level 386
 Constant odds-ratio (COR) model, *see* Odds ratio, constant
 Construction of probability models, *see* Assessment; Strategies, assessment; Model building, statistical
 Convergence
 pointwise 79
 of relative frequencies, *see* Frequency, relative
 of upper and lower probabilities 3–4, 219, 220, 223–25, 470–71, 620–21
 Convex combination 79, 93
 Convex hull 611, 613
 Convexity 611
 of set of coherent lower previsions 79
 of set of dominating linear previsions 145–46
 of set of gambles 162
 Countable additivity, *see* Additivity, countable
 Credible intervals 383, 387
 Currency, probability (*see also* Utility) 25, 59
 Daniell integral 493
 Data, *see* Observation; Information
 Death risk 352, 354–55, 360
 Decision 24–26, 160–66, 235–41
 Bayesian theories 43, 164–65, 237–38
 experimental studies 48, 116–17
 frequentist theory 164–65
 fuzzy analysis 263
 with imprecise utilities 165–66
 with precise utilities 160–65
 role of probability models in 24–25, 161
 sensitivity analysis 24, 44, 238, 256
 statistical 164–65
 strategies for 24–25, 238–41
 versus inference 21, 109–10, 572
 see also Action

Decision rules
 Bayes 162, 164
 minimax regret 508
 P -minimax 163–64, 240, 269–70, 619, 630
 statistical 164–65, 240
 Decisiveness 235–36, 243
 Definition, operational 20, 102–03, 243, 247, 624
 Dempster-Shafer theory, *see* Belief functions
 Dempster's rule of combination 275–81
 Dempster's rule of conditioning 278–81
 Density function (*see also* Distribution)
 conditional 331
 conjugate, *see* Density function, natural-conjugate
 Haldane 223, 228, 374–76, 417–18
 improper 228, 229–30, 369
 behavioural meaning 230–31
 generates incoherent inferences 231, 369–77
 objections to 230–35, 376–77
 relation to proper previsions 233, 376, 399, 417–18, 420–21
 uniform 229–35, 369–72, 376, 383–84, 399, 420–21
 joint 331–32, 450
 marginal 331, 393
 natural-conjugate 205
 noninformative 226–35
 for a chance 228–29
 dependence on sampling model 229, 231–32, 374–76, 583
 for a location parameter 229–34, 370–72, 376, 420–21
 for location-scale parameters 232, 373, 383
 motivations for 234–35
 objections to 230–34
 as a reference prior 233–34, 583
 for a scale parameter 372–73
see also Density function, improper posterior 393
 regular 437–39
 sampling 369, 393, 436–37
 uniform, *see* Density function, improper, uniform; Distribution, uniform

upper and lower 200–02
 Dependence (*see also* Independence)
 degree of 589
 non-negative 473
 Desirability 150–53, 155, 156–60, 614–16
 almost- 151–53
 axioms for, *refer to* Index of axioms B- 287, 294
 behavioural interpretation 60, 151, 614
 correspondence with conditional previsions 615–16
 correspondence with linear previsions 158
 correspondence with lower previsions 64, 156, 615
 correspondence with preference 153
 more fundamental than prevision 159–60
 real 160, 614–16
 role in elicitation 152, 169, 171–72
 strict 155
 Determinacy (*see also* Indeterminacy) 209
 Determinism 354–55, 358–59
 Direct inference, *see* Principles of rationality, direct inference
 Disagreement, *see* Conflict
 Dispositions
 behavioural 18–20, 61–63
 physical, *see* Chance
 to update beliefs 285–86
see also Beliefs
 Distribution
 beta 205, 218, 395
 binomial 374–75, 395
 degenerate 149, 209
 double exponential 384
 geometric 375–76
 maximum entropy 266–72
 noninformative, *see* Density function, noninformative
 Normal 370–72
 imprecise 398–99, 399–400, 401–03, 405, 423–24, 524, 620
 with unknown variance 355–57, 373, 382–83
 posterior, *see* Prevision, posterior;

Density function, posterior prior, *see* Prevision, prior; Density function, prior
 Student's t 373, 382–83
 uniform
 on bounded interval 372, 384–85
 on finite set 227–28, 267
 on integers 310
 on positive integers 97–98, 143–44, 311, 321–22, 326, 365, 450
 on real line 99–100, 136, 144, 420–21, *see also* Density function, improper, uniform
 on surface of a sphere 328
 on unit interval 98–99, 220, 223, 228, 437–39, *see also* Lebesgue measure
 Distribution function 130
 upper and lower 130, 203–05
 Domain
 of conditional previsions 289, 291, 293
 of lower prevision 61
 of statistical model 363, 388, 393
 of upper prevision 70
 Dominance 78, 132–35
 Drawing pin, *see* Thumbtack
 Dutch book 29, *see also* Loss
 Economics, examples of difference between buying and selling prices 65, 72, 94–95, 95–97, 602–07, 627–28
 Elicitation 167–74
 Bayesian procedures 173–74
 of beliefs about football game 172–73, 175–77, 632–33
 general procedure 167–76
 incomplete 104, 106, 173, 215–17
 operational procedures 173, 622–31
 using structural judgements 474–76
 versus assessment 14–15, 167
see also Assessment; Judgement; Measurement of probabilities
 Embedding principle, *see* Principles of rationality, embedding
 Entropy 211, 266
 for continuous distributions 271
 upper and lower 541
 Envelopes
 of Bayesian models 397–400, 402–03, 405, 418–19, 429–30, 432
 of coherent linear collections 308, 312–13, 316, 327, 349, 414–15, 642–44
 of countably additive measures 327, 454
 independent lower 446–48, 454–57
 of linear previsions 89, 132–36
 relation to coherence 134–35, 145
 relation to natural extension 136, 414
 of lower previsions 78, 348–49
 of noninformative priors 223–24
 of variances 618
see also Sensitivity analysis
 Equivalence judgements 472
 Equivalence of models for uncertainty 156, 159–60
 Estimation
 interval, *see* Confidence intervals; Credible intervals
 point 619
 Events 54, 81
 conditioning 284, 289
 of probability zero 306–07, 328–34, 436–40
 exchangeable 460–67
 independent 443–48
 irrelevant 444
 non-negatively dependent 473
 non-negatively relevant 473
 observable 118, 465–66
 permutable 458–59
 probable 188–91, 262
 standard 196
 sure 54
 as zero-one valued gambles 81
see also Gamble
 Evidence, *see* Information
 Exchangeability 361, 460–67
 as a direct judgement of desirability 472
 grounds for 461–62, 466
 for infinite sequence of events 463–66
 stronger than permutability 461–62
 Expectation (*see also* Prevision) 128–29, 493

Experiments
 auxiliary 196
 binary, *see* Bernoulli models
 exchangeable 461
 independent 448
 permutable 457
see also Trials
 Expert systems 49–50
 for assessing probabilities 40, 512
 Explanation
 for actions, psychological 17–19
 versus prediction 360, 466
 Extension
 of conditional and marginal
 previsions 314–17, 413
 to conditional previsions 317–33,
 412, 615–16, 639–41
 from a field 125, 127–32
 Hahn–Banach 138–39
 of independent marginals 452–57
 to joint previsions 314–17, 405, 415–
 24, 452–57
 of Lebesgue measure, *see* Lebesgue
 measure
 to linear previsions 136–39, 418–21,
 427–28, 433
 existence of 137, 418
 uniqueness of 129, 138, 419–20,
 428
 from a linear space 125–27
 maximal coherent 126
 minimal coherent (*see also* Extension,
 natural) 123–24, 313–14, 318,
 408–15, 425, 430
 natural 30–31, 121–32, 408–15
 basic properties 123, 409–10
 behavioural meaning 122, 408–09
 in betting on horses 606–07
 of classificatory judgements 189
 of comparative probability
 ordering 192
 computation 128, 130, 136, 174–76,
 316
 of desirable gambles 151–53
 via generalized Bayes rule 318,
 425–27, 429, 431–32
 independent 453–57
 as lower envelope of linear

extensions 136, 316, 414, 418,
 429, 432
 relation to coherence 74, 122–23,
 408–11
 role in assessment and
 elicitation 35–36, 171–72
 role in statistical reasoning 406
 of structural judgements 475
 to posterior previsions 424–34,
 640–41
 existence of 425–27, 430–31
 for imprecise sampling
 models 430–34, 440–41
 as lower envelope of Bayesian
 posteriors 429, 432
 satisfies likelihood principle 427,
 431–32
 uniqueness of 425, 428
 to predictive previsions 314, 421–24
 to prior previsions 415–24
 existence of 416, 422
 as lower envelope of Bayesian
 priors 418
 typically non-vacuous 368, 417
 uniqueness of 419–20, 423
 regular 582, 616, 639–41
 to unconditional previsions (*see also*
 Extension, to joint previsions;
 Extension, to prior
 previsions) 121–32, 314–17, 413
 Extreme point 146, 613
 Extreme points of $\mathcal{M}(P)$ 146–48, 159
 computation 175, 511
 examples 147
 existence 146, 148
 lower envelopes of 147, 149, 429, 432
 Family, location, *see* Parameter, location
 Field 90
 extension from 125, 127–32
 σ - 126
 Filter 100–01, 131, 147–49, 190, 310
 de Finetti's elimination of aleatory
 probabilities 111–13, 361, 465–
 67
 de Finetti's fundamental theorem of
 probability 138
 de Finetti's representation theorem 463

Finite additivity, *see* Additivity
 Football
 examples of assessments 172, 175,
 179, 180, 181, 182, 185, 187, 190,
 194, 198, 199, 201, 202, 270, 288,
 634
 World Cup experiment 632–38
 Frequency, relative 219, 350–52, 357–58
 apparent convergence 464–65
 convergence 351, 354, 464
 divergence 80–81, 358–59, 469
 see also Interpretations of probability
 and precision, frequency
 Fubini's theorem 588
 Function, *see* Belief function; Density
 function; etc.
 Functional
 bivariate 613, 618
 evaluation 609
 linear 609
 continuous 609, 611–13
 positive 89, 609
 variance 618
 Fuzzy sets 50, 261–66
 Gain, sure 64
 avoiding 87
 from betting on horses 603–05
 of bookmakers 95, 627–28
 in football experiment 635, 637–38
 in the stock market 72
 Gamble 58
 admissible 161
 \mathcal{A} -measurable 129
 Bayes 162
 \mathcal{B} -measurable 291
 called off 56, 184–85, 284
 contingent 284
 desirable, *see* Desirability
 indeterminate 474, 476
 marginal 68, 289
 maximal 161
 nondesirable 474
 permuted 457
 P -minimax 163
 simple 129, 174
 super-relevant 381
 two-stage 291, 294, 303–04
 undesirable 474
 Gambling, *see* Betting
 Generalized Bayes rule (GBR), *see* Bayes
 rule, generalized
 Geometric representation of probability
 models 176–77, 494, 512
 Geometry, Euclidean 249–50
 Hahn–Banach extensions 138–39
 Hahn–Banach theorem 139, 506, 611
 Haldane density, *see* Density function,
 Haldane
 Horse racing, betting on 94–95, 602–07
 Hyperparameter 260
 Hyperplane
 separating 133, 611
 simplex representation for 176
 strictly separating 611–12
 strongly separating 612
 supporting 146
 Hyperplane theorems 611–12
 Hyperprior 260
 Hypotheses, statistical, *see* Models,
 sampling
 Hypothesis testing, examples of 233,
 583
 Idealization
 in model building 39, 117
 of precise measurement 329, 436–40
 of precise probabilities 249–50
 Identifiability 350, 471, 563
 Ignorance 226–35
 about a chance 228–29
 about a finite space 227–28
 about a location parameter 229–34
 about a positive integer 322
 Bayesian models for 223–24, 226–35
 complete 92–93, 226–27, 270, 445,
 458, 462
 near- 206, 218, 235
 beta model for 218–21, 621, 641
 partial, *see* Indeterminacy
 posterior 366–68
 consistent with precise prior 310,
 391–92
 consistent with prior
 ignorance 308, 391

Ignorance (*Contd.*)
 prior 367, 417
 implies posterior ignorance 367–69, 428–29
see also Prevision, vacuous
Imprecision in probabilities 210
 arguments against 235–53
 arguments for 3–8, 212–17
 degree of 210, 219, 224–26, 468
 in football experiment 633–34
 as measure of information 211–12, 222
 due to conflict between information 213, 222–26
 due to lack of information 212–13, 217–22, 224–26
 models for 50–51, 159–60, 253–81
 in second-order probabilities 260–61
 sources of 212–17
see also Indeterminacy; Precision (of probabilities)
Improper prior, *see* Density function, improper
Inadmissibility, *see* Admissibility
Incoherence
 examples of 29, 67, 72, 85–86, 198–99, 300, 601, 605, 636
 of standard statistical models 370, 372, 373–76, 382–83
see also Loss, examples of
Incomparability, *see* Completeness
Incompleteness of probability
 models 104–05, 173, 212, 215–17
Inconsistency, *see* Conflict
Indecision 226, 235–41
 versus suspension of judgement 239, 530
 ways of resolving 238–41
Independence 443–57
 behavioural meaning 443, 448
 conditional 451
 epistemic 443–44, 448
 of events 443–48
 of experiments 448–57
 grounds for 444–45
 linear 596
 logical 445
 physical 444

SUBJECT INDEX

sensitivity analysis definitions 446–48, 454–57
 standard definitions 445–46, 448–51
 as a structural judgement 473
Indeterminacy 209–10
 in conditional revisions 306–07, 328–29
 experimental evidence for 48, 116–17, 254, 532, 633
 judgements of 474, 476
 in likelihood function 436
 may produce indecision 161–62, 236
 physical 215, 357–58, 468
 in posterior revisions 395–96, 436–40
 sources of 212–15
 versus incompleteness 104–06, 212
see also Imprecision in probabilities
Indicator function 81
Indifference, principle of, *see* Principle of indifference
Induction, *see* Logic, inductive
Inference, *see* Reasoning; Inference, statistical
Inference, statistical 21–22, 406, 424
 Bayesian 5, 42–43, 393–97
 arguments against 3–8, 110–11, 114, 226–35, 397
 arguments for 252–53
 objective 43, 226–35, 266–72, 369–77
 robust (*see also* Sensitivity analysis) 5–7, 44, 256–58
 standard 393–97
 subjective 43
 behavioural theories 21–23, 109–10
 objections to 109–119
fiducial 373, 382
frequentist (*see also* Confidence intervals) 377–88
 arguments against 109–10, 114, 252–53, 387–88
 from imprecise priors 217–26, 397–403, 424–34
 from improper priors 369–77, 417–18, 420–21
 from noninformative priors 226–35, 369–77, 383

SUBJECT INDEX

likelihood 377
 arguments against 109–10, 114, 377
 arguments for 434–37, 568
see also Likelihood principle
 limits to precision 215, 358–59, 471
Neyman–Pearson, *see* Confidence intervals
objective methods 367, 373–74, 377
 examples of incoherence 369–76, 379, 382–85
 not consistent with prior ignorance 368–69, 417–18, 420–21
pivotal 374
 requires non-vacuous prior beliefs 110, 367–69, 387, 391, 417, 429
structural 373
 versus decision 21, 109
INFERO 50
Infimum 58
Information 33
 amount of 222
 comparison of measures 211–12, 268, 523–24
 reflected in degree of imprecision 3–4, 211, 220, 222
 relation to sample size 213, 222
 assessment of, *see* Assessment; Strategies, assessment
 from Bernoulli trials 217–26
 combination of, *see* Combination of assessments
 conflicting 187–88, 213–14, 223–26
 incomplete, *see* Information, partial
 of limited relevance 214, 221
 needed to determine conditional revisions 329–34, 437–40
new 336–38
 can increase indeterminacy 223–26, 298–300, 433, 551
 modelled as an event 300–01, 336–39
 versus old 338
partial 212–13
 prior 110, 114, 221–22
 relevant 33, 444
Shannon's theory of 268, 523–24
 statistical 434–35
 symmetric 33, 227, 458, 461–62
see also Observation; Ignorance
Initial conditions 355, 358–59
Instability
 in elicitation 216–17
 physical 215, 358–59
 of relative frequencies 359, 469
Insurance premiums 95, 245, 532
Integral
 constructed from measure 128–29
 Daniell 493
 Lebesgue 132, 493
Intentional systems 17–18
Interpretations of probability and
 precision 13–17, 101–108
 aleatory (*see also* frequency or propensity) 14, 16–17, 349–59
 relation to epistemic 17, 102, 353–54, 357
 of sampling models 111–14, 349–59, 465–66
 behavioural 14–24, 61–63, 109–10
 minimal 20–21, 61–63
 objections to 109–19
 versus evidential 22–24
 versus sensitivity analysis 105–08, 476–77
 constructive 14–16
 contingent 284–85
 descriptive 361
 direct 105, 357
 dispositional 15, 16, 18–20
 empirical, *see* aleatory
 epistemic 14–17, 102
 needed in inference 15, 100
 relation to aleatory 17, 353–54, 357
 of sampling models 360–61
 epistemological, *see* logical
 evidential 14, 22–24
 exhaustive 62, 104–05, 251, 476
 frequency 16, 49
 finite 350–51, 357–58
 limiting 49, 80–81, 351, 358, 378
 incomplete 104–05, 285
 instrumentalist 360–61

SUBJECT INDEX

Interpretations (*Contd.*)
 intersubjective 361
 logical 14–16, 43, 102, 267
 in decision making 238
 in inductive logic 34–35, 44–45,
 48–49
 as version of sensitivity analysis 107, 259
 objective, *see aleatory or logical observational* 14–16
 operational 15, 20, 102–03
 personalist 14, 43
 allows irrational judgements 32,
 110, 253
 in decision making 237–38
 as version of sensitivity analysis 107
 propensity 16–17, 351–55, 358
 rationalistic 14–15, 102
 reductionist 112–13, 361, 465–67
 sensitivity analysis (*see also Sensitivity analysis*) 7, 105–08, 253–58
 based on dogma of ideal precision 7, 105, 254
 and conglomerability 312–13, 327
 of elicited probabilities 173, 477
 of structural judgements 108, 446–48, 456–57, 460, 477
 versus behavioural interpretation 107–08, 477
 subjective, *see personalist*
 theoretical 15–16, 103
 updated 284–86
 Interval estimation, *see Confidence intervals; Credible intervals*
 Interval of measures 201–02
 Invariance 139–45, 231
 permutation 457
 shift 97–98, 143–44
 translation 98–100, 140, 144–45, 229–30
 Irrelevance 444, *see also Independence*
 Jeffrey's rule 339
 Judgement (*see also Assessment; Elicitation*) 167–71
 ambiguous versus imprecise 263
 arbitrary 173–74, 210, 237, 241

SUBJECT INDEX

classificatory 188–91
 comparative 191–97
 of desirability 169–70, 474
 effect of new 177–80
 equivalence 472
 structural 472–77
 suspending 226, 239
 types of 169–71, 472–74

Keynes's theory of probability 44–45
 Kolmogorov's condition 306
 Kolmogorov's theory of conditional probability 306–08
 compared to de Finetti's theory 282, 327
 Krein–Milman theorem 613

Laplace's rule of succession 220
 Laws, deterministic versus statistical 352, 355
 Laws of large numbers 352
 for imprecise probabilities 359, 564
 Lebesgue lower prevision 98–99, 132
 is fully conglomerable 325
 Lebesgue measure (on real line) 229–30,
see also Density function, improper, uniform
 Lebesgue measure (on unit interval) 98–99
 additive extensions 132, 136–37, 144–45, 325–27
 inner and outer 98–99, 131–32, 136–37, 149–50, 325, 327
 non-measurable sets 149–50
 translation-invariant extensions 144–45

Likelihood function 400, 427
 for continuous sample space 435–37
 imprecise 431
 medial 431–32
 upper and lower 431, 441
 yields vacuous inferences (without prior assessments) 366–69

Likelihood inference, *see Inference, statistical, likelihood*

Likelihood principle 434–41
 for continuous sample spaces 435–40
 for discrete sample spaces 434–35

for imprecise likelihoods 440–41
 satisfied by natural extension 427, 431–32
 satisfied by regular extension 641
 violated by frequentist inferences 583
 violated by noninformative priors 229, 232, 583

Likelihood ratios, upper and lower 433

Linearity 66, 86–89, 609
 sub- 65
 super- 63
see also Prevision, linear; Additivity

Linear programming 175, 511

Linear space, *see Space, linear*

Linear-vacuous mixture, *see Mixture, linear-vacuous*

Linkage between prior and sampling model 399–400, 405, 423–24

Location, *see Parameter, location*

Logic
 deductive 485, 494
 fuzzy 50, 265–66, 538
 inductive (*see also Interpretations of probability and prevision, logical*) 34–35, 44–45

Loss function, *see Utility*

Loss, in non-statistical problems 28–29, 67–72, 114–16
 axioms for 68, 293–94, 302–03
 examples of 28, 67, 72, 85, 177, 187, 279, 280, 288, 311, 312, 321–23, 604, 635–36
see also Avoiding sure loss

Loss, in statistical problems 343–46
 partial 344–45
 axioms for 344, 363, 380, 389, 576
 examples of (*see also other types of loss*) 375–76

sure 346
 axioms for 346, 364, 380, 389
 examples of 370, 372, 373, 375, 379, 382–83, 384, 390

uniform 346
 axioms for 346, 364, 380, 389
 examples of 365–66, 385

uniform sure 343–44
 axioms for 343, 364–65, 380, 389, 390

examples of 365, 369–70, 371–72, 379, 385, 585

Lottery tickets, as unit of utility, *see Currency, probability*

Mapping, multivalued 182–84, 273–74

Marginalization 181–82

Marginals 181–82, 314
 compatible 453
 independent 448
see also Prevision, joint

Market economics 65, 72, 493, 627–28

Maximality in preference ordering 161
 relation to Bayesian optimality 162–63, 165–66

Mean, invariant 140

Measurability
 Lebesgue 98, 149–50
 with respect to a field 129, 306
 with respect to a partition 291, 306

Measure (*see also Additivity, countable*)
 construction of integral from 128–29
 inner and outer 126
 intervals of, *see Interval of measures*
 Lebesgue, *see Lebesgue measure*
 probability, *see Additivity, countable product* 453–54

Measurement
 biased, model for 399–400, 405, 423–24
 imprecise 250, 329, 436–40
see also Observation

Measurement of probabilities, procedures for 173, 622–31
 biases in 624–28, 631
 bookmaking procedure 625–28
 for comparing assessors 630, 632–35
 comparison with standard events 196–97
 general elicitation procedure 168–74
 multiple-price procedure 628–29, 630
see also Scoring rules
 scoring rules, *see Scoring rules*
 symmetric procedure 629–33
 two-sided betting procedure 243, 624–25
see also Elicitation

Mechanics, Newtonian 249–50

Medical diagnosis 40, 50, 529

Membership function 261–66
 assessment of 262, 264–65
 Zadeh's rules of combination 265–66

Methods of statistical inference, *see*
Inference, statistical

Minimax decision, *see* Decision rules;
Action, minimax

Minimax theorem 613

Mixture, *see* Convex combination

Mixture, linear-vacuous 93–94, 135,
 147, 202–03
 preserved under conditioning 309,
 311–12, 325

Model building, statistical 35–37, 38,
 359–60, 467–68, *see also*
Strategies, assessment

Models
Bayesian, see Models, standard Bayesian
Bernoulli, see Bernoulli models
constant odds-ratio, see Odds ratio, constant
finitely generated 174–75
hierarchical 37, 232, 259–60
neighbourhood, see Neighbourhood of a linear revision
pari-mutuel, see Betting, pari-mutuel system
sampling 349–61
approximate 355–56
construction of 359–60, 467–68
continuous 435–40
imprecise 356–60, 399–400, 405, 423–24, 430–34, 440–41, 467–71
interpretations of, see Interpretations of probability and revision
*robust 356–57, *see also* Models, sampling, imprecise*
standard Bayesian 393–97, 397–400, 405
statistical 341–43
Money pump 28
Monotonicity
complete 272
of preferences 154, 156
*of previsions 76, 88
 2– 130*
MYCIN 50

SUBJECT INDEX

Natural extension, *see* Extension, natural
 Near-ignorance, *see* Ignorance, near-
 Neighbourhood of a linear revision
 comparison of models 96–97, 202–03
 constant odds-ratio, *see* Odds ratio, constant
 linear-vacuous mixture, *see* Mixture, linear-vacuous
 pari-mutuel, *see* Betting, pari-mutuel system

Neyman–Pearson theory, *see* Confidence intervals

Noninformative prior, *see* Density function, noninformative

Norm, supremum 609

Normal distribution, *see* Distribution, Normal

Norms of rationality, *see* Principles of rationality

Notation, *refer to* Glossary of notation 674

Nuclear risk 246

Objectivity
 different meanings of 111–14
 of statistical methods, *see* Inference, statistical, objective methods

Observability 118–19, 350, 465–66, 624

Observation 284, 336–37
 continuous 328–32, 436–40
 modelled as an event 284, 336–38
 unexpected 336, 525
 unreliable 274, 276, 339

Odds, upper and lower 96, 433, 603, *see also* Betting rates

Odds ratio, constant (COR model) 95–97, 135, 147, 202–03, 401–03
 preserved under conditioning 309–10
 preserved under statistical updating 401

Operationalism, *see* Definition, operational

Opinions, *see* Beliefs

Orderings
 comparative probability, *see* Probability, comparative

preference, *see* Preference

SUBJECT INDEX

Paradox of ideal evidence 220

Parameter 349
 learning 219
 location 229, 372, 383–84
 location-scale 232, 373
 meaningful 350, 465–66
 nuisance 357, 468
 region of values 205, 218
 scale 372
 true value 350, 355–56, 360

Parameter space 349, 362

Pari-mutuel model, *see* Betting, pari-mutuel system

Partition of possible observations 284, 289

Penalty function, *see* Scoring rules

Permutability 457–60
 compared to exchangeability 460–62
 and convergence of relative frequencies 465
 examples of 459, 469
 grounds for 458
 as a structural judgement 473, 475–76

Permutation-invariance, *see* Invariance

Physicality 111–13

Possibility 54–58
 apparent 55
 epistemic 55
 new 184–85
 practical 56
 pragmatic 56, 185

Possibility space 54
 in decision problems 24, 239
 modifications of 56–57, 180–86, 337–38
 observable 118
 in statistical problems 362

Posterior revision, *see* Revision, posterior

Precision of measurement, *see* Measurement, imprecise

Precision (of probabilities) 210
 arguments against 3–8, 43–44, 226–35, 397
 arguments for 235–53
 axioms of 31–32, 91–92, 241–48
 Bayesian dogma of 3, 241
 dogma of ideal 7, 105–07, 254–57

frequentist dogma of 564
 ideal versus idealization 106–07, 249, 254
 limits to 215, 216, 359, 471
 not necessary for decision 236, 243
see also Imprecision in probabilities; Revision, linear

Prediction, probabilistic 422
 versus explanation 360, 466

Preference 47, 153–56
 almost- 153–54
 axioms for, *refer to* Index of axioms 680
 behavioural meaning 153, 236–237
 complete 157, 241–47
 correspondence with desirability 153
 correspondence with linear revisions 158
 correspondence with lower revisions 156
 intransitive 28–29
 more fundamental than revision 159
 real 616
 role in decision making 160–62
 strict 155–56, 161
 versus choice 236–37, 242, 247

Revision (*see also* Probability; Interpretations of probability and revision)
 conditional 282–340
 axioms for, *refer to* Index of axioms 680
 coherence with unconditional previsions 293–94, 301–05
 defined through generalized Bayes rule 297
 not determined by unconditional 306–07, 310, 328–29, 450–51
 on events of probability zero 328–34
 examples of 298–300, 308–13, 321–22, 325
 linear 304–06
 roles in updating and assessment 285–87, 334–40
 vacuous 293, 298, 308, 313, 328
see also Revision, posterior;

Prevision (*Contd.*)

 Prevision, updated;
 Conglomerability
 contingent (*see also* Prevision,
 conditional) 284–89
 joint
 constructed from conditional and
 marginal previsions 313–17,
 421–24
 constructed from independent
 marginals 452–57
 constructed from permutable
 marginals 459, 475–76
 not determined by conditional and
 marginal previsions 404–05,
 423–24
 exchangeable 460–61
 independent (*see also* Prevision,
 product) 452
 permutable 457
 linear 66, 86–92
 axioms for, *refer to* Index of axioms
 680
 non-conglomerable 311, 312, 321–
 23
 determined by additive
 probability 91, 128–29
 envelopes of, *see* Envelopes
 extensions of 136–39
 ideal 105–07, 254–57, 258–59
 invariant 139–45
 true, *see* ideal
 see also Probability, additive;
 Precision (of probabilities)
 lower 61
 axioms for, *refer to* Index of axioms
 680
 basic properties 76–80
 behavioural interpretation 61–63
 conglomerable 317
 non-conglomerable 311
 correspondence with other
 models 145, 147, 156
 not determined by lower
 probability 82–84, 177
 examples of 92–101
 interpretations of (*see also*
 Interpretations of probability and
 prevision) 101–08
 as lower envelope of linear

SUBJECT INDEX

 previsions 134, 145–47
 translation-invariant 98–100, 421
 see also Probability, lower
 marginal 181–82, 314
 posterior 362, 424
 coherence with prior
 ignorance 367–69
 coherence with prior
 previsions 388–405, 416, 418
 coherence with sampling
 model 362–67
 defined through generalized Bayes
 rule 400–03, 427, 432
 not determined by prior
 assessments 337–38, 395–96,
 399–400, 433–34, 437–40
 may be less precise than prior
 precision 223–26, 298–300,
 433, 551
 needed to measure uncertainty in
 conclusions 22, 110, 252, 388
 regular 640–41
 regular Bayesian 438–39
 standard Bayesian 393–96, 428
 vacuous 298, 308, 366–67, 391–92,
 428–29, 438
 see also Prevision, conditional
 predictive 422
 prior 388
 assessment of, *see* Assessment, of
 prior previsions
 Bayesian 393–97, 418–20
 conjugate 205–06
 data-dependent 336, 338, 525, 583–
 84
 implied by sampling model and
 posterior previsions 368–69, 387,
 415–21, 429
 imprecise 397–403
 improper, *see* Density function,
 improper
 joint 403–05
 maximum entropy 266–72
 near-ignorance 206, 218–21
 needed in statistical inference 110,
 368–69, 428–29
 noninformative, *see* Density
 function, noninformative
 vacuous 391, 417, 419, 468
 product 452–57

SUBJECT INDEX

 linear 453
 types 1 and 2 455–57
 second-order, *see* Probability, second-
 order
 unconditional, *see* Prevision, lower;
 Prevision, upper; Prevision, linear
 updated 284–89, 334–39, *see also*
 Prevision, posterior; Prevision,
 conditional; Updating
 upper 64–65, 70, 75
 basic properties 76–77
 behavioural interpretation 64–65
 see also Prevision, lower
 vacuous 66, 92–93, 308
 equivalent models 135, 147, 157
 as model for complete
 ignorance 92–93, 228
 versus uniform distributions 93, 228
 Price for gamble
 buying 61, 625, 629–30
 contingent 284
 updated 284, 303, 305
 fair (buying and selling) 66, 86, 241–
 44
 selling 65, 625, 629
 see also Betting rates; Prevision, linear;
 Prevision, lower
 Principle of direct inference, *see*
 Principles of rationality, direct
 inference
 Principle of indifference 227, 267
 Principle of insufficient reason, *see*
 Principle of indifference
 Principle of maximum entropy
 (PME) 266–72
 Principles of rationality 27–32, 33–35
 avoiding sure loss, *see* Avoiding sure
 loss; Loss
 coherence, *see* Coherence
 conglomerability, *see* Conglomerability
 conglomerative 294–96, 304–05, 552
 contingent version 295–96, 345
 direct inference 17, 33, 353–55
 embedding 227–28
 ignorance 34, 206, 212–13
 independence 33–34, 444–45
 likelihood, *see* Likelihood principle
 natural extension, *see* Extension,
 natural
 symmetry 33, 227, 458

 updating 287–89, 304–05, 550

Prior prevision, *see* Prevision, prior

Prisoners problem 279, 299–300

Probability (*see also* Prevision;
 Interpretations of probability and
 prevision)

additive (*see also* Additivity; Prevision,
 linear) 89–91

correspondence with linear
 prevision 91, 128–29

zero-one valued 100–01, 147–50

classificatory 188–91, 197, 248

comparative 45–46, 191–97, 241,
 244–46, 600–01

comparison with other scientific
 concepts 249–50

conditional, *see* Prevision, conditional

countably additive, *see* Additivity,
 countable

currency 25, 59

density function, *see* Density function

distribution, *see* Distribution

fuzzy 262

imprecise (*see also* Imprecision in
 probabilities, Indeterminacy) 51

literature on 44–50

types of 50–51, 156, 159

interval-valued, *see* Probability, lower

lower 28, 46–47, 81–86, 197–99
 axioms for, *refer to* Index of axioms
 680

completely monotone 272

does not determine lower
 precision 82–84, 177

as lower envelope of additive
 probabilities 135

n-coherent 600

2-monotone 130

zero-one valued 100–01, 131, 136,
 147–49, 310, 313

see also Prevision, lower

measure, *see* Additivity, countable

non-additive, *see* Probability, lower

posterior, *see* Prevision, posterior

precise, *see* precision (of probabilities);
 Prevision, linear; Probability,
 additive

prior, *see* Prevision, prior

relation to evidence 23, 32–42

second-order 258–61, 264

Probability (*Contd.*)
 examples of 259, 271
 imprecise 260–61
 unconditional, *see* Probability, lower
 upper (*see also* Probability,
 lower) 82–85, 197–99
 as a betting rate 82, 602–03
 vacuous, *see* Prevision, vacuous
Product
 of previsions, *see* Prevision, product
 of sets, *see* Space, product
Propensity (*see also* Chance) 351–55
Propositions 491
Psychology of action 17–20
Psychology, experimental studies of
 decision and indeterminacy 48,
 116–17, 254, 632–34
Quantiles, upper and lower 204
Quantum mechanics 358, 480, 491
Radioactivity 352
Radon-Nikodym derivative 436, 574
Randomization
 in decision making 162
 in experimental design 299
Randomness (*see also* Chance) 358
 versus ignorance 113–14, 212
Random quantity 57–58, *see also*
 Gamble
Random variable, *see* Gamble
Rates, betting, *see* Betting rates
Rates, tax, *see* Taxation, models for
Rationality 26–42
 Aristotelean 484
 bounded 40–41, 599
 external 27–28, 33–42, 113
 internal 27–32, 42
 principles of, *see* Principles of
 rationality
 versus reasoning 26
see also Reasoning; Principles of
 rationality
Reason, limitations of 40–41, 116–17, 241
Reasoning 26–27, 30–31, 35–42
 inductive, *see* Reasoning, probabilistic;
 Logic, inductive
 principles of, *see* Principles of
 rationality
 probabilistic 1, 21–22

SUBJECT INDEX

standards of 26–27
 statistical 406
 versus rationality 26
see also Rationality; Inference,
 statistical; Extension, natural;
 Strategies, assessment
Reductionism, *see* Interpretations of
 probability and prevision,
 reductionist
Refinement (of a possibility space) 56–
 57, 180–81, 337
Regret 508
Relevance, *see* Information, relevant
Relevant subset 381–86
Reward
 from action 160
 in elicitation 624, 629–31
Risk function 164
Robustness 5–6
 of Bayesian inferences 6, 256, *see also*
 Sensitivity analysis
 of Bernoulli models 468
 of independence judgements 589
 of inferences from conjugate
 priors 206
 of likelihood function 431
 of sampling models 356–57
Roles of probability-utility models 17
 constructive 21–26, 30–31, 35–42
 in decision 24–25, 160–61
 explanatory (descriptive) 17–19, 116–
 17
 in inference 21–22, 26, 30–31
 normative 27–35, 41–42
Rules, *see* Decision rules; Scoring rules;
 Updating rules; Bayes' rule;
 Conjunction rule; Unanimity rule;
 etc.
Saddle point 613, 619
Sample equivalent to prior
 information 221–22
Sample space 362
 continuous 328–33, 435–40
see also Possibility space
Sampling models, *see* Models, sampling
Satisfaction level 240
Satisficing 239–40
Scores 630
Scoring rules 247–48, 630–31

SUBJECT INDEX

absolute error 248
 deterministic 534
 logarithmic 269, 631
 as operational definition of
 prevision 247, 533–34
 proper 631
 quadratic 247, 270, 631
Self-conjugacy 66, 87, 90
Semigroup of transformations 139
 Abelian 139, 142–45
Sensitivity analysis 6–7, 44, 105–08,
 171, 253–58
 Berger's account 256–57
 in decision problems 238
 Good's 'black box' model 255–56
 models for independence 446–48,
 454–57
 models for permutation-
 invariance 460–61
 practical differences from behavioural
 theory 107–08, 477
 similarities to behavioural
 theory 107–08, 257–58
 in statistical inference 257–58, 397–
 400
Separating hyperplane theorems 611–12
 strength of 505–06
Sequences of previsions 79–81, 326
Sets
 compact, *see* Compactness
 convex, *see* Convexity
 fuzzy, *see* Fuzzy sets
 Lebesgue-measurable 98
 non-measurable 149–50
Sex 481
 σ -field, *see* Field, σ -; Borel σ -field
Simplex, probability 176–77
 examples of 83, 176, 179, 183, 186,
 188, 191, 194, 198, 199, 201, 202,
 635
Space
 dual 609
 finite dimensional 609, 612
 half 146, 610
 linear 63, 608
 locally convex 612
 parameter, *see* Parameter space
 possibility, *see* Possibility space
 product 196, 330–33, 362, 448
 sample, *see* Sample space
topological 609
vector, *see* Space, linear
Stake 59, 69
 in betting on football games 633–34,
 637
 effective 630
Stakes function 629, 630–31
Standard deviation, upper and
 lower 617
 dependence on sample size 620–21
see also Variance
State of affairs 54
 observable versus theoretical 54–55,
 118
 personal interest in 623, 624
 possible, *see* Possibility
Statistics, theories of, *see* Inference,
 statistical
Stimulus-response theory 19–20
Strategies
 assessment 8, 35–42, 286
 conflicting 214
 criteria for selection 38–40
 as justification for assessments
 41–42
 in statistical problems 406
 types of 37–38
see also Assessment; Extension
 decision-making 24–25, 160–66, 238–
 41
 inference 21–22, 406, 424
 updating (*see also* Updating) 285–86,
 335–40, 406, 424
Student's t-distribution, *see* Distribution,
 Student's t
Subjectivity, *see* Objectivity
Subspace, linear 608
Support of gamble 344, 363
Supremum 58
Symmetry, in information 227, 458,
 461–62
Symmetry principles, *see* Principles of
 rationality, symmetry
Systems, expert, *see* Expert systems
Systems, intentional, *see* Intentional
 systems
Taxation, models for 96, 604
 capital gains 95–96
t-distribution, *see* Distribution, Student's t

SUBJECT INDEX

- Theorems; *see* Minimax theorem, Separating hyperplane theorems; etc.
- Thumbtack (drawing pin), probability models 36–37
- for imprecise chances 433–34, 454–55, 467–68
- for near-ignorance 4, 218–20
- objectivity of 112–13, 465–66
- for prior-data conflict 225–26
- robust Bernoulli model 467–71
- standard Bernoulli model 217, 467, 470
- for uncertain perception 339
- Topology 609–10
- supremum-norm 609
- weak* 145–46, 609–10
- Tractability of models 105, 116–17, 216, 252
- Transitivity of preferences 28–29, 154, 247
- Translation-invariance, *see* Invariance
- Trials 351
- Bernoulli, *see* Bernoulli models
- repeated 351–54, 358, 359, 462–64, 467
- single 351–52, 354
- see also* Experiments
- Two-monotonicity, *see* Monotonicity, 2-
- Ultrafilter 101, 147–49
- Ultrafilter theorem 148, 149
- Unanimity rule 188, 214
- Uncertainty 209
- degree of 211–12, 266, 617, *see also* Entropy; Variance
- determinate, *see* Determinacy
- indeterminate, *see* Indeterminacy
- mathematical models for 51, 156, 159
- choice between 159–60
- equivalence of 156
- literature on 42–50
- need for posterior measures 22, 110, 252, 388
- types of 113–14
- see also* Beliefs; Indeterminacy; Revision; Probability
- Uniform distribution, *see* Distribution, uniform; Density function, improper, uniform
- Updating 285–86, 334–40
- coherent 317–19, 321–22
- difficulties in 334–37
- a general strategy 337–38
- statistical 336, 406, 424
- see also* Conditioning; Bayes rule, generalized; Revision, conditional; Revision, posterior; Conglomerability
- Updating principle, *see* Principles of rationality, updating
- Updating rules 285–86, 297, *see also* Bayes' rule; Bayes rule, generalized; Dempster's rule of conditioning; Jeffrey's rule
- Updating strategies, *see* Strategies, updating
- Urn, drawings from 462–63
- Utility 17–19, 24–26, 58–61
- construction of scale 59–60
- expected 18, 24, 162
- imprecise 165–66, 490
- precise (linear) 25–26, 59–61, 64, 115, 160
- quadratic 58, 231, 619
- unbounded 58, 527
- Utility function, identified with gamble 25, 160–61
- Vacuous precision, *see* Precision, vacuous
- Vagueness 245–46, 261, 519, *see also* Ambiguity; Indeterminacy
- Values 17–19, *see also* Utility
- Variable, continuous 330–33, 436–40
- Variable, random, *see* Gamble
- Variance of a gamble
- precise 617, 619–20
- upper and lower 617–21
- Variance of a sample 373
- Verification procedure 54, 118
- approximate 119
- see also* Definition, operational
- Well-groundedness (of probability assessments) 35, 113–14
- Williams' theorem 643
- Witness, unreliable 274, 276
- You 18