

# Decision Theory

## Principles and Approaches

**Giovanni Parmigiani**

*Johns Hopkins University, Baltimore, USA*

**Lurdes Y. T. Inoue**

*University of Washington, Seattle, USA*

*with contributions by*

**Hedibert F. Lopes**

*University of Chicago, USA*



A John Wiley and Sons, Ltd., Publication



# Decision Theory

## **WILEY SERIES IN PROBABILITY AND STATISTICS**

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

### **Editors**

DAVID J. BALDING, NOEL A. C. CRESSIE, GARRETT M. FITZMAURICE, IAIN M. JOHNSTONE, GEERT MOLENBERGHS, DAVID W. SCOTT, ADRIAN F. M. SMITH, RUEY S. TSAY, SANFORD WEISBERG, HARVEY GOLDSTEIN.

### **Editors Emeriti**

VIC BARNETT, J. STUART HUNTER, JOZEF L. TEUGELS

A COMPLETE LIST OF THE TITLES IN THIS SERIES APPEARS AT THE END OF THIS VOLUME.

# Decision Theory

## Principles and Approaches

**Giovanni Parmigiani**

*Johns Hopkins University, Baltimore, USA*

**Lurdes Y. T. Inoue**

*University of Washington, Seattle, USA*

*with contributions by*

**Hedibert F. Lopes**

*University of Chicago, USA*



A John Wiley and Sons, Ltd., Publication

This edition first published 2009  
© 2009 John Wiley & Sons, Ltd.

*Registered office*

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex,  
PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at [www.wiley.com](http://www.wiley.com).

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

***Library of Congress Cataloging-in-Publication Data***

Parmigiani, G. (Giovanni)

Decision theory : principles and approaches / Giovanni Parmigiani, Lurdes Inoue.  
p. cm.

Includes bibliographical references and index.

ISBN 978-0-471-49657-1 (cloth)

1. Statistical decision. 2. Axiomatic set theory. 3. Experimental design.

I. Inoue, Lurdes. II. Title.

QA279.4.P37 2009

519.5'42—dc22

2009008345

A catalogue record for this book is available from the British Library.

ISBN 978-0-471-49657-1

Set in 10/12pt Times by Integra Software Services Pvt. Ltd, Pondicherry, India  
Printed in the UK by TJ International, Padstow, Cornwall

To our advisors: Don Berry, Morrie De Groot, and Jay Kadane;  
and to their advisors: Jay Kadane, Jimmy Savage, and Herman Chernoff





# Contents

<b>Preface</b>	<b>xiii</b>
<b>Acknowledgments</b>	<b>xvii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Controversies	1
1.2 A guided tour of decision theory	6
<b>Part One Foundations</b>	<b>11</b>
<b>2 Coherence</b>	<b>13</b>
2.1 The “Dutch Book” theorem	15
2.1.1 <i>Betting odds</i>	15
2.1.2 <i>Coherence and the axioms of probability</i>	17
2.1.3 <i>Coherent conditional probabilities</i>	20
2.1.4 <i>The implications of Dutch Book theorems</i>	21
2.2 Temporal coherence	24
2.3 Scoring rules and the axioms of probabilities	26
2.4 Exercises	27
<b>3 Utility</b>	<b>33</b>
3.1 St. Petersburg paradox	34
3.2 Expected utility theory and the theory of means	37
3.2.1 <i>Utility and means</i>	37
3.2.2 <i>Associative means</i>	38
3.2.3 <i>Functional means</i>	39
3.3 The expected utility principle	40
3.4 The von Neumann–Morgenstern representation theorem	42
3.4.1 <i>Axioms</i>	42
3.4.2 <i>Representation of preferences via expected utility</i>	44
3.5 Allais’ criticism	48
3.6 Extensions	50
3.7 Exercises	50

<b>4</b>	<b>Utility in action</b>	<b>55</b>
4.1	The “standard gamble”	56
4.2	Utility of money	57
4.2.1	<i>Certainty equivalents</i>	57
4.2.2	<i>Risk aversion</i>	57
4.2.3	<i>A measure of risk aversion</i>	60
4.3	Utility functions for medical decisions	63
4.3.1	<i>Length and quality of life</i>	63
4.3.2	<i>Standard gamble for health states</i>	64
4.3.3	<i>The time trade-off methods</i>	64
4.3.4	<i>Relation between QALYs and utilities</i>	65
4.3.5	<i>Utilities for time in ill health</i>	66
4.3.6	<i>Difficulties in assessing utility</i>	69
4.4	Exercises	70
<b>5</b>	<b>Ramsey and Savage</b>	<b>75</b>
5.1	Ramsey’s theory	76
5.2	Savage’s theory	81
5.2.1	<i>Notation and overview</i>	81
5.2.2	<i>The sure thing principle</i>	82
5.2.3	<i>Conditional and a posteriori preferences</i>	85
5.2.4	<i>Subjective probability</i>	85
5.2.5	<i>Utility and expected utility</i>	90
5.3	Allais revisited	91
5.4	Ellsberg paradox	92
5.5	Exercises	93
<b>6</b>	<b>State independence</b>	<b>97</b>
6.1	Horse lotteries	98
6.2	State-dependent utilities	100
6.3	State-independent utilities	101
6.4	Anscombe–Aumann representation theorem	103
6.5	Exercises	105
<b>Part Two</b>	<b>Statistical Decision Theory</b>	<b>109</b>
<b>7</b>	<b>Decision functions</b>	<b>111</b>
7.1	Basic concepts	112
7.1.1	<i>The loss function</i>	112
7.1.2	<i>Minimax</i>	114
7.1.3	<i>Expected utility principle</i>	116
7.1.4	<i>Illustrations</i>	117
7.2	Data-based decisions	120
7.2.1	<i>Risk</i>	120
7.2.2	<i>Optimality principles</i>	121

7.2.3	<i>Rationality principles and the Likelihood Principle</i>	123
7.2.4	<i>Nuisance parameters</i>	125
7.3	The travel insurance example	126
7.4	Randomized decision rules	131
7.5	Classification and hypothesis tests	133
7.5.1	<i>Hypothesis testing</i>	133
7.5.2	<i>Multiple hypothesis testing</i>	136
7.5.3	<i>Classification</i>	139
7.6	Estimation	140
7.6.1	<i>Point estimation</i>	140
7.6.2	<i>Interval inference</i>	143
7.7	Minimax–Bayes connections	144
7.8	Exercises	150
<b>8</b>	<b>Admissibility</b>	<b>155</b>
8.1	Admissibility and completeness	156
8.2	Admissibility and minimax	158
8.3	Admissibility and Bayes	159
8.3.1	<i>Proper Bayes rules</i>	159
8.3.2	<i>Generalized Bayes rules</i>	160
8.4	Complete classes	164
8.4.1	<i>Completeness and Bayes</i>	164
8.4.2	<i>Sufficiency and the Rao–Blackwell inequality</i>	165
8.4.3	<i>The Neyman–Pearson lemma</i>	167
8.5	Using the same $\alpha$ level across studies with different sample sizes is inadmissible	168
8.6	Exercises	171
<b>9</b>	<b>Shrinkage</b>	<b>175</b>
9.1	The Stein effect	176
9.2	Geometric and empirical Bayes heuristics	179
9.2.1	<i>Is <math>x</math> too big for <math>\theta</math>?</i>	179
9.2.2	<i>Empirical Bayes shrinkage</i>	181
9.3	General shrinkage functions	183
9.3.1	<i>Unbiased estimation of the risk of <math>x + g(x)</math></i>	183
9.3.2	<i>Bayes and minimax shrinkage</i>	185
9.4	Shrinkage with different likelihood and losses	188
9.5	Exercises	188
<b>10</b>	<b>Scoring rules</b>	<b>191</b>
10.1	Betting and forecasting	192
10.2	Scoring rules	193
10.2.1	<i>Definition</i>	193
10.2.2	<i>Proper scoring rules</i>	194
10.2.3	<i>The quadratic scoring rules</i>	195
10.2.4	<i>Scoring rules that are not proper</i>	196

10.3	Local scoring rules	197
10.4	Calibration and refinement	200
	10.4.1 <i>The well-calibrated forecaster</i>	200
	10.4.2 <i>Are Bayesians well calibrated?</i>	205
10.5	Exercises	207
<b>11</b>	<b>Choosing models</b>	<b>209</b>
11.1	The “true model” perspective	210
	11.1.1 <i>Model probabilities</i>	210
	11.1.2 <i>Model selection and Bayes factors</i>	212
	11.1.3 <i>Model averaging for prediction and selection</i>	213
11.2	Model elaborations	216
11.3	Exercises	219
<b>Part Three</b>	<b>Optimal Design</b>	<b>221</b>
<b>12</b>	<b>Dynamic programming</b>	<b>223</b>
12.1	History	224
12.2	The travel insurance example revisited	226
12.3	Dynamic programming	230
	12.3.1 <i>Two-stage finite decision problems</i>	230
	12.3.2 <i>More than two stages</i>	233
12.4	Trading off immediate gains and information	235
	12.4.1 <i>The secretary problem</i>	235
	12.4.2 <i>The prophet inequality</i>	239
12.5	Sequential clinical trials	241
	12.5.1 <i>Two-armed bandit problems</i>	241
	12.5.2 <i>Adaptive designs for binary outcomes</i>	242
12.6	Variable selection in multiple regression	245
12.7	Computing	248
12.8	Exercises	251
<b>13</b>	<b>Changes in utility as information</b>	<b>255</b>
13.1	Measuring the value of information	256
	13.1.1 <i>The value function</i>	256
	13.1.2 <i>Information from a perfect experiment</i>	258
	13.1.3 <i>Information from a statistical experiment</i>	259
	13.1.4 <i>The distribution of information</i>	264
13.2	Examples	265
	13.2.1 <i>Tasting grapes</i>	265
	13.2.2 <i>Medical testing</i>	266
	13.2.3 <i>Hypothesis testing</i>	273
13.3	Lindley information	276
	13.3.1 <i>Definition</i>	276
	13.3.2 <i>Properties</i>	278
	13.3.3 <i>Computing</i>	280
	13.3.4 <i>Optimal design</i>	281

13.4	Minimax and the value of information	283
13.5	Exercises	285
<b>14</b>	<b>Sample size</b>	<b>289</b>
14.1	Decision-theoretic approaches to sample size	290
14.1.1	<i>Sample size and power</i>	290
14.1.2	<i>Sample size as a decision problem</i>	290
14.1.3	<i>Bayes and minimax optimal sample size</i>	292
14.1.4	<i>A minimax paradox</i>	293
14.1.5	<i>Goal sampling</i>	295
14.2	Computing	298
14.3	Examples	302
14.3.1	<i>Point estimation with quadratic loss</i>	302
14.3.2	<i>Composite hypothesis testing</i>	304
14.3.3	<i>A two-action problem with linear utility</i>	306
14.3.4	<i>Lindley information for exponential data</i>	309
14.3.5	<i>Multicenter clinical trials</i>	311
14.4	Exercises	316
<b>15</b>	<b>Stopping</b>	<b>323</b>
15.1	Historical note	324
15.2	A motivating example	326
15.3	Bayesian optimal stopping	328
15.3.1	<i>Notation</i>	328
15.3.2	<i>Bayes sequential procedure</i>	329
15.3.3	<i>Bayes truncated procedure</i>	330
15.4	Examples	332
15.4.1	<i>Hypotheses testing</i>	332
15.4.2	<i>An example with equivalence between sequential and fixed sample size designs</i>	336
15.5	Sequential sampling to reduce uncertainty	337
15.6	The stopping rule principle	339
15.6.1	<i>Stopping rules and the Likelihood Principle</i>	339
15.6.2	<i>Sampling to a foregone conclusion</i>	340
15.7	Exercises	342
	<b>Appendix</b>	<b>345</b>
A.1	Notation	345
A.2	Relations	349
A.3	Probability (density) functions of some distributions	350
A.4	Conjugate updating	350
	<b>References</b>	<b>353</b>
	<b>Index</b>	<b>367</b>



# Preface

## Goals

The goal of this book is to give an overview of fundamental ideas and results about rational decision making under uncertainty, highlighting the implications of these results for the philosophy and practice of statistics. The book grew from lecture notes from graduate courses taught at the Institute of Statistics and Decision Sciences at Duke University, at the Johns Hopkins University, and at the University of Washington. It is designed primarily for graduate students in statistics and biostatistics, both at the Masters and PhD level. However, the interdisciplinary nature of the material should make it interesting to students and researchers in economics (choice theory, econometrics), engineering (signal processing, risk analysis), computer science (pattern recognition, artificial intelligence), and scientists who are interested in the general principles of experimental design and analysis.

Rational decision making has been a chief area of investigation in a number of disciplines, in some cases for centuries. Several of the contributions and viewpoints are relevant to both the education of a well-rounded statistician and to the development of sound statistical practices. Because of the wealth of important ideas, and the pressure from competing needs in current statistical curricula, our first course in decision theory aims for breadth rather than depth. We paid special attention to two aspects: bridging the gaps among the different fields that have contributed to rational decision making; and presenting ideas in a unified framework and notation while respecting and highlighting the different and sometimes conflicting perspectives.

With this in mind, we felt that a standard textbook format would be too constraining for us and not sufficiently stimulating for the students. So our approach has been to write a “tour guide” to some of the ideas and papers that have contributed to making decision theory so fascinating and important. We selected a set of exciting papers and book chapters, and developed a self-contained lecture around each one. Some lectures are close to the source, while others stray far from their original inspiration. Naturally, many important articles have been left out of the tour. Our goal was to select a set that would work well together in conveying an overall view of the fields and controversies.

We decided to cover three areas: the axiomatic foundations of decision theory; statistical decision theory; and optimal design of experiments. At many universities,

these are the subject of separate courses, often taught in different departments and schools. Current curricula in statistics and biostatistics are increasingly emphasizing interdisciplinary training, reflecting similar trends in research. Our plan reflects this need. We also hope to contribute to increased interaction among the disciplines by training students to appreciate the differences and similarities among the approaches.

We designed our tour of decision-theoretic ideas so that students might emerge with their own overall philosophy of decision making and statistics. Ideally that philosophy will be the result of contact with some of the key ideas and controversies in the different fields. We attempted to put the contributions of each article into some historical perspective and to highlight developments that followed. We also developed a consistent unified notation for the entire material and emphasized the relationships among different disciplines and points of view. Most lectures include current-day materials, methods, and results, and try at the same time to preserve the viewpoint and flavor of the original contributions.

With few exceptions, the mathematical level of the book is basic. Advanced calculus and intermediate statistical inference are useful prerequisites, but an enterprising student can profit from most of the the book even without this background. The challenging aspect of the book lies in the swift pace at which each lecture introduces new and different concepts and points of view.

Some lectures have grown beyond the size that can be delivered during a  $1\frac{1}{2}$  hour session. Some others merge materials that were often taught as two separate lectures. But for the most part, the lecture–session correspondence should work reasonably well. The style is also closer to that of transcribed lecture notes than that of a treatise. Each lecture is completed by worked examples and exercises that have been helpful to us in teaching this material. Many proofs, easy and hard, are left to the student.

## Acknowledgments

We have intellectual debt to more people than we can list, but a special place in this list is occupied by courses we took and lecture notes we read. Giovanni’s course at Duke was initially developed from two main sources. The first is the lectures from Teddy Seidenfeld’s course on the Foundations of Statistics. Giovanni only took it five times—he will firmly hold he did not choose the stopping rule: left to his own devices he would have taken that class forever. The second is the lectures from Schervish’s course on Advanced Statistics, from which his book on the theory of statistics would ultimately develop. We also had access to a very insightful bootleg of Charles Stein’s lecture notes at Stanford, from an edition of the course taught by Persi Diaconis.

Dennis Lindley reviewed an early draft and gave very constructive comments and encouragement. Other anonymous reviewers gave helpful feedback. Bruno Sansó used our notes to teach his class at the University of California at Santa Cruz, and gave us detailed comments. We have used our notes in teaching for over a decade. Many students braved earlier drafts, gave useful feedback through questions, conversations, solutions to problems, and sometimes highly informative puzzled looks. Martin McIntosh shared his precious correspondence with Herman Chernoff.



Both of us are grateful to Hedibert Lopes, with whom our long journey to writing this book had started back in the mid 1990s. His notes from Giovanni's classes were used extensively in early versions of this book, and some figures, problems, and examples still carry his hallmark.

Lurdes is thankful to Sergio Wechsler who opened the door to new ways of thinking about statistics and introducing her to decision theory. She thanks Giovanni for inviting her on this journey, which through bumps and laughter has been a lifetime experience. She cannot wait for the next one (well, give and take some time off for her recovery from the thrill!). She wishes to thank the loving support from her brothers Roberto, Carlos, and Silvio and from her uncles Masao, Olinda, and Tadazumi. Finally, her loving gratitude goes to her parents, Satie and Kasuo, and her grandmother Matta, for the inspiring memories and lessons that guide Lurdes.

Giovanni still has mixed feelings about the day Marco Scarsini handed him a copy of Wald's book on decision functions, with the assignment of reporting on it to an undergraduate discussion group. Later Michele Cifarelli, Guido Consonni, Morrie DeGroot, Jay Kadane, Teddy Seidenfeld, Mark Schervish, Nick Polson, Don Berry, Pietro Muliere, Peter Müller, and David Matchar fueled his intellectual passion for rational decision making. Giovanni's wife Francesca is a statistician who, despite her impact on national policy making at various levels, is a bit bored by the kind of decision theory her husband favors, and perhaps baffled by the scant influence all the talking about rationality has had on his personal behavior. Nevertheless, she has been fully supportive of this never-ending project, in more ways than one can list. Giovanni thinks working with Lurdes has been absolutely fantastic. He has not told her yet, but he is already thinking about notation changes for the second edition . . .

Giovanni Parmigiani, Lurdes Y. T. Inoue



# Acknowledgments

1. Page 15; Extracted from Savage LJ (1981a). A panel discussion of personal probability, *The writings of Leonard Jimmie Savage – A memorial Selection*, American Statistical Association, Alexandria, VA, pp. 508–513, American Statistical Association.
2. Pages 16, 22; Extracted from de Finetti, B. (1937). Foresight: its logical laws, its subjective sources, in H. E. Kyburg and H. E. Smokler (eds.), *Studies in Subjective Probability*, Krieger, New York, pp. 55–118.
3. Pages 24–25; Extracted from Goldstein, M. (1983). The prevision of a prevision, *Journal of the American Statistical Association* 78: 817–819, American Statistical Association.
4. Pages 36–37, 48, 82–83, 88, 115–116; Extracted from Savage, L. J. (1954). *The foundations of statistics*, John Wiley & Sons, Inc., New York.
5. Page 38; Extracted from Bonferroni, C. (1924). La media esponenziale in matematica finanziaria, *Annuario del Regio Istituto Superiore di Scienze Economiche e Commerciali di Bari*, Vol. 23–24, pp. 1–14, Regio Istituto Superiore di Scienze Economiche e Commerciali di Bari.
6. Pages 61, 63; Extracted from Pratt, J. (1964). Risk aversion in the small and in the large, *Econometrica* 32: 122–136, The Econometric Society.
7. Page 93; Extracted from Ellsberg, D. (1961). Risk, ambiguity and the savage axioms, *Quarterly Journal of Economics* 75: 643–669, The MIT Press.
8. Page 93; Extracted from Gärdenfors, P. and Sahlin, N.-E. (1988). *Decision, Probability and Utility*, Cambridge University Press, Cambridge.
9. Pages 97–98; Extracted from Savage, L. J. (1981a). A panel discussion of personal probability, *The writings of Leonard Jimmie Savage – A memorial selection*, American Statistical Association, Alexandria, VA, pp. 508–513, American Statistical Association.
10. Page 99; Extracted from Anscombe, F. J. and Aumann, R. J. (1963). A definition of subjective probability, *Annals of Mathematical Statistics* 34: 199–205, Institute of Mathematical Statistics.
11. Page 103; Extracted from Schervish, M. J., Seidenfeld, T. and Kadane, J. B. (1990). State-dependent utilities, *Journal of the American Statistical Association* 85: 840–847, American Statistical Association.
12. Page 111; Extracted from Edgeworth, F. Y. (1887). The method of measuring probability and utility, *Mind* 12 (47): 484–488, Oxford University Press.

13. Page 113; Extracted from Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses, *Annals of Mathematical Statistics* 10: 299–326, Institute of Mathematical Statistics.
14. Page 115; Extracted from Chernoff, H. (1954). Rational selection of a decision function, *Econometrica* 22: 422–443, The Econometric Society.
15. Page 133; Extracted from Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient test of statistical hypotheses, *Philosophical Transaction of the Royal Society (Series A)* 231: 286–337, Springer-Verlag, New York.
16. Pages 150–151; Extracted from Lindley, D. V. (1968b). Decision making, *The Statistician* 18: 313–326, Blackwell Publishing.
17. Pages 153–154; Extracted from Birnbaum, A. (1962). On the foundations of statistical inference (Com: P307-326), *Journal of the American Statistical Association* 57: 269–306. (From Pratt's comments to Birnbaum 1962, pp. 314–315), American Statistical Association.
18. Pages 114, 283; Extracted from Savage, L. J. (1951). The theory of statistical decision, *Journal of the American Statistical Association* 46: 55–67, American Statistical Association.
19. Page 171; Extracted from Schervish, M. J. (1995). *Theory of Statistics*, Springer-Verlag, American Statistical Association.
20. Page 175; Extracted from Robert, C. P. (1994). *The Bayesian Choice*, Springer-Verlag.
21. Pages 191–192; Extracted from Brier, G. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review* 78: 1–3, American Meteorological Society.
22. Page 199; Extracted from Good, I. J. (1952). Rational decisions, *Journal of the Royal Statistical Society, Series B, Methodological* 14: 107–114, Blackwell Publishing.
23. Page 200; Extracted from DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters, *The Statistician* 32: 12–22, Blackwell Publishing.
24. Page 207; Extracted from Dawid, A. (1982). The well-calibrated Bayesian, *Journal of the American Statistical Association* 77: 605–613, American Statistical Association.
25. Page 224; Extracted from Nemhauser, G. (1966). *Introduction to Dynamic Programming*, John Wiley & Sons, Inc., New York.
26. Page 225; Extracted from Lindley, D. (1961). *Dynamic programming and decision*, *Applied Statistics*, Blackwell Publishing.
27. Pages 226, 228–229; Extracted from Lindley, D. V. (1985). *Making Decisions*, second ed., John Wiley & Sons, Ltd, Chichester.
28. Pages 251–253; Extracted from French, S. (1988). *Decision theory: an introduction to the mathematics of rationality*, Ellis Horwood, Chichester, Horwood Publishing Ltd.
29. Page 291; Extracted from Grundy, P., Healy, M. and Rees, D. (1956). Economic choice of the amount of experimentation, *Journal of The Royal Statistical Society. Series B.* 18: 32–55, Blackwell Publishing.

30. Pages 297–298; Extracted from Lee, S. and Zelen, M. (2000). Clinical trials and sample size considerations: another perspective, *Statistical Science* 15: 95–110, Institute of Mathematical Statistics.
31. Page 324–325; Extracted from Wald, A. (1945). Sequential tests of statistical hypotheses, *Annals of Mathematical Statistics* 16: 117–186, Institute of Mathematical Statistics.
32. Page 324; Extracted from Wallis, W. A. (1980). The Statistical Research Group, 1942-1945 (C/R: p331–335), *Journal of the American Statistical Association* 75: 320–330, American Statistical Association.
33. Page 339; Extracted from Berger, J. and Berry, D. (1988). The relevance of stopping rules in statistical inference (with discussion), in J. Berger and S. Gupta (eds.), *Statistical Decision Theory and Related Topics IV*, Vol. 1, Springer-Verlag, New York, pp. 29–72.
34. Page 341; Extracted from Kadane, J. B., Schervish, M. J. and Seidenfeld, T. (1996). Reasoning to a foregone conclusion, *Journal of the American Statistical Association* 91: 1228–1235, American Statistical Association.



# 1

## Introduction

We statisticians, with our specific concern for uncertainty, are even more liable than other practical men to encounter philosophy, whether we like it or not.

(Savage 1981a)

### 1.1 Controversies

Statistics is a mature field but within it remain important controversies. These controversies stem from profoundly different perspectives on the meaning of learning and scientific inquiry, and often result in widely different ways of interpreting the same empirical observations.

For example, a controversy that is still very much alive involves how to evaluate the reliability of a prediction or guess. This is, of course, a fundamental issue for statistics, and has implications across a variety of practical activities. Many are captured by a case study on the evaluation of evidence from clinical trials (Ware 1989). We introduce the controversy with an example. You have to guess a secret number. You know it is an integer. You can perform an experiment that would yield either the number before it or the number after it, with equal probability. You know there is no ambiguity about the experimental result or about the experimental answer. You perform this type of experiment twice and get numbers 41 and 43. What is the secret number? Easy, it is 42. Now, how good an answer do you think this is? Are you tempted to say “It is a perfect answer, the secret number has to be 42”? It turns

out that not all statisticians think this is so easy. There are at least two opposed perspectives on how to go about figuring out how good our answer is:

**Judge an answer by what it says.** *Compare the answer to other possible answers, in the light of the experimental evidence you have collected, which can now be taken as a given.*

versus

**Judge an answer by how it was obtained.** *Specify the rule that led you to give the answer you gave. Compare your answer to the answers that your rule would have produced when faced with all possible alternative experimental results. How far your rule is from the truth in this collection of hypothetical answers will inform you about how good your rule is. Indirectly, this will tell you how well you can trust your specific answer.*

Let us go back to the secret number. From the first perspective you would compare the answer “42” to all possible alternative answers, realize that it is the only answer that is not ruled out by the observed data, and conclude that there is no ambiguity about the answer being right. From the second perspective, you ask how the answer was obtained. Let us consider a reasonable recipe for producing answers: average the two experimental results. This approach gets the correct answer half the time (when the two experimental results differ) and is 1 unit off the remainder of the time (when the two experiments yield the same number). Most measures of error that consider this entire collection of potential outcomes will result in a conclusion that will attribute some uncertainty to your reported answer. This is in sharp contrast with the conclusion reached following from the first perspective. For example, the standard error of your answer is  $1/\sqrt{2}$ . By this principle, you would write a paper reporting your discovery that “the secret number is 42 (s.e. 0.7)” irrespective of whether your data are 41 and 43, or 43 and 43. You can think of other recipes, but if they are to give you a single guess, they are all prone to making mistakes when the two experimental results are the same, and so the story will have the same flavor.

The reasons why this controversy exists are complicated and fascinating. When things are not as clear cut as in our example, and multiple answers are compatible with the experimental evidence, the first perspective requires weighing them in some way—a step that often involves judgment calls. On the other hand the second perspective only requires knowing the probabilities involved in describing how the experiments relate to the secret number. For this reason, the second approach is perceived by many to be more objective, and more appropriate for scientific inquiry. Objectivity, its essence, worthiness, and achievability, have been among the most divisive issues in statistics. In an extreme simplification the controversy can be captured by two views of probability:

**Probability lives in the world.** *Probability is a physical property like mass or wavelength. We can use it to describe stochastic experimental*



*mechanisms, generally repeatable ones, like the assignments of experimental units to different conditions, or the measurement error of a device. These are the only sorts of probabilistic considerations that should enter scientific investigations.*

versus

**Probability lives in the mind.** *Probability, like most conceptual constructs in science, lives in the context of the system of values and theories of an individual scientist. There is no reason why its use should be restricted to repeatable physical events. Probability can for example be applied to scientific hypotheses, or the prediction of one-time events.*

Ramsey (1926) prefaced his fundamental paper on subjective probability with a quote from poet William Blake: “Truth can never be told so as to be understood, and not be believed.”

These attitudes define a coordinate in the space of statisticians’ personal philosophies and opinions, just like the poles of the previous controversy did. These two coordinates are not the same. For example, there are approaches to the secret number problem that give different answers depending on whether data are 41 and 43, or 43 and 43, but do not make use of “subjective” probability to weigh alternative answers. Conversely, it is common practice to evaluate answers obtained from subjective approaches, by considering how the same approaches would have fared in other experiments.

A key aspect that both these dimensions have in common is the use of a stochastic model as the basis for learning from data. In the secret number story, for example, the starting point was that the experimental results would fall to the left or right of the secret number *with equal probability*. The origin of the role of probability in interpreting experimental results is sampling. The archetype of many statistical theories is that experimental units are sampled from a larger population, and the goal of statistical inference is to draw conclusions about the whole population. A *statistical model* describes the stochastic mechanism based on which samples are selected from the population. Sometimes this is literally the case, but more often samples and populations are only metaphors to guide the construction of statistical procedures. While this has been the model of operation postulated in most statistical theory, in practical applications it is only one pole of yet another important controversy:

**Learning requires models.** *To rigorously interpret data we need to understand and specify the stochastic mechanism that generated them. The archetype of statistical inference is the sample-population situation.*

versus

**Learning requires algorithms.** *To efficiently learn from data, it is critical to have practical tools for exploring, summarizing, visualizing,*

*clustering, classifying. These tools can be built with or without explicit consideration of a stochastic data-generating model.*

The model-based approach has ancient roots. One of the relatively recent landmarks is Fisher's definition of the likelihood function (Fisher 1925). The algorithmic approach also goes back a long way in history: for example, most measures of dependence, such as the correlation coefficient, were born as descriptive, not inferential tools (Galton 1888). The increasing size and complexity of data, and the interface with computing, have stimulated much exploratory data analysis (Tukey 1977, Chambers *et al.* 1983) and statistical work at the interface with artificial intelligence (Nakhaeizadeh and Taylor 1997, Hastie *et al.* 2003). This controversy is well summarized in an article by Breiman (2001).

The premise of this book is that it is useful to think about these controversies, as well as others that are more technical in statistics, from first principles. The principles we will bring to bear are principles of rationality in action. Of course, this idea is in itself controversial. With this regard, the views of many statisticians distribute along another important dimension of controversy:

**Statisticians produce knowledge.** The scope of statistics is to rigorously interpret experimental results, and present experimental evidence in an unbiased way to scientists, policy makers, the public, or whoever may be in charge of drawing conclusions or making decisions.

versus

**Statisticians produce solutions to problems.** Understanding data requires placing them in the context of scientific theories, which allow us to sort important from ancillary information. One cannot answer the question "what is important?" without first considering the question "important for what?"

Naturally, producing knowledge helps solving problems, so these two positions are not in contrast from this standpoint. The controversy is on the extent to which the goals of an experiment should affect the learning approaches, and more broadly whether they should be part of our definition of learning.

The best known incarnation of this controversy is the debate between Fisher and Neyman about the meaning of hypothesis tests (Fienberg 1992). The Neyman–Fisher controversy is broader, but one of the key divides is that the Neyman and Pearson theory of hypothesis testing considers both the hypothesis of interest and at least one alternative, and involves an explicit quantification of the consequences of rejecting or accepting the hypothesis based on the data: the type I and type II errors. Ultimately, Neyman and Pearson's theory of hypothesis testing will be one of the key elements in the development of formal approaches to rationality-based statistical analysis. On the other hand Fisher's theory of significance test does not require considering an alternative and incarnates a view of science in which hypotheses represent working

approximations to natural laws, that serve to guide experimentation, until they are refuted with sufficient strength that new theories evolve.

A simple example challenges theories of inference that are based solely on the evidence provided by observation, regardless of scope and context of the theory. Let  $x_1$  and  $x_2$  be two Bernoulli trials. Suppose the experimenter's probabilities are such that  $P(x_1 = 0) = P(x_2 = 0) = 0.5$  and  $P(x_1 + x_2 = 0) = 0.05$ . Then,  $P(x_1 + x_2 = 1) = 0.9$  and  $P(x_1 + x_2 = 2) = 0.05$ . Let  $e$  be the new evidence that  $x_1 = 1$ , let  $h_1$  be the hypothesis that  $x_1 + x_2 = 2$ , and  $h_2$  be the hypothesis that  $x_2 = 1$ . Given  $e$ , the two hypotheses are equivalent. Yet, probability-wise,  $h_1$  is corroborated by the data, whereas  $h_2$  is not. So if one is to consider the change in probability as a measure of support for a theory, one would be left with either an inconsistent measure of evidence, or the need to defend the position that the two hypotheses are in some sense different even when faced with evidence that proves that they are the same. This and other similar examples seriously question the idea that inductive practice can be adequately represented by probabilities alone, without relation to their rational use in action.

There can be disagreements of principle about whether consideration of consequences and beliefs belongs to scientific inquiry. In reality, though, it is our observation that the vast majority of statistical inference approaches have an implicit or explicit set of goals and values that guide the various steps of the construction. When making a decision as simple as summarizing a set of numbers by their median (as opposed to, say, their mean) one is making judgments about the relative importance of the possible oversimplifications involved. These could be made formal, and in fact there are decision problems for which each of the two summaries is optimal. Our view is that scientific discussion is more productive when goals are laid out in the open, and perhaps formalized, than when they are hidden or unappreciated. As the old saying goes, "there are two types of statisticians: those who know what decision problem they are solving and those who don't."

Despite the draconian simplifications we have made in defining the dimensions along which these four controversies unfold, one would be hard pressed to find two statisticians that live on the same point in this four-dimensional space. One of the goals of this book is to help students find their own spot in a way that reflects their personal intellectual values, and serves them best in approaching the theoretical and applied problems that are important to them.

We definitely lean on the side of "judging answers by what they say" and believing that "probabilities live in the mind." We may be some distance away along the models versus algorithm dimension—at least judging by our approaches in applications. But we are both enthusiastic about the value of thinking about rationality as a guide, though sometimes admittedly a rough guide, to science, policy, and individual action. This guidance comes at two levels: it tells us how to formally connect the tools of an analysis with the goals of that analysis; and it tells us how to use rationality-based criteria to evaluate alternative statistical tools, approaches, and philosophies.

Overall, our book is an invitation to Bayesian decision-theoretic ideas. While we do not think they necessarily provide a solution to every statistical problem, we

find much to think about in this comment from Herman Chernoff (from a personal communication to Martin McIntosh):

Frankly, I am not a Bayesian. I go under the following principle. If you don't understand a problem from a Bayesian decision theory point of view, you don't understand the problem and trying to solve it is like shooting at a target in the dark. Once you understand the problem, it is not necessary to attack it from a Bayesian point of view. Bayesian methods always had the difficulty that our approximations to our subjective beliefs could carry a lot more information than we thought or felt willing to assume.

## 1.2 A guided tour of decision theory

We think of this book as a “tour guide” to the key ideas in decision theory. The book grew out of graduate courses where we selected a set of exciting papers and book chapters, and developed a self-contained lecture around each one. We make no attempt at being comprehensive: our goal is to give you a tour that conveys an overall view of the fields and its controversies, and whets your appetite for more. Like the small pictures of great paintings that are distributed on leaflets at the entrance of a museum, our chapters may do little justice to the masterpiece but will hopefully entice you to enter, and could guide you to the good places.

As you read it, keep in mind this thought from R. A. Fisher:

The History of Science has suffered greatly from the use by teachers of second-hand material, and the consequent obliteration of the circumstances and the intellectual atmosphere in which the great discoveries of the past were made. A first-hand study is always useful, and often . . . full of surprises. (Fisher 1965)

Our tour includes three parts: foundations (axioms of rationality); optimal data analysis (statistical decision theory); and optimal experimental design.

**Coherence.** We start with de Finetti’s “Dutch Book Theorem” (de Finetti 1937) which provides a justification for the axioms of probability that is based on a simple and appealing rationality requirement called coherence. This work is the mathematical foundation of the “probabilities live in the mind” perspective. One of the implications is that new information is merged with the old via Bayes’ formula, which gets promoted to the role of a universal inference rule—or Bayesian inference.

**Utility.** We introduce the axiomatic theory of utility, a theory on how to choose among actions whose consequences are uncertain. A rational decision maker proceeds by assigning numerical utilities to consequences, and scoring actions by their expected utility. We first visit the birthplace of quantitative utility: Daniel Bernoulli’s St. Petersburg paradox (Bernoulli 1738). We then present in detail von Neumann and

Morgenstern's utility theory (von Neumann and Morgenstern 1944) and look at a criticism by Allais (1953).

**Utility in action.** We make a quick detour to take a look at practical matters of implementation of rational decision making in applied situations, and talk about how to measure the utility of money (Pratt 1964) and the utility of being in good health. For applications in health, we examine a general article (Torrance *et al.* 1972) and a medical article that has pioneered the utility approach in health, and set the standard for many similar analyses (McNeil *et al.* 1981).

**Ramsey and Savage.** We give a brief digest of the beautiful and imposing axiomatic system developed by Savage. We begin by tracing its roots to the work of Ramsey (1931) and then cover Chapters 2, 3, and 5 from Savage's *Foundations of statistics* (Savage 1954). Savage's theory integrates the coherence story with the utility story, to create a more general theory of individual decision making. When applied to statistical practice, this theory is the foundation of the "statisticians find solutions to problems" perspective. The general solution is to maximize expected utility, and expectations are computed by assigning personal probabilities to all unknowns. A corollary is that "answers are judged by what they say."

**State independence.** Savage's theory relies on the ability to separate judgment of probability from judgment of utility in evaluating the worthiness of actions. Here we study an alternative axiomatic justification of the use of subjective expected utility in decision making, due to Anscombe and Anmann (1963). Their theory highlights very nicely the conditions for this separation to take place. This is the last chapter on foundations.

**Decision functions.** We visit the birthplace of statistical decision theory: Wald's definition of a general statistical decision function (Wald 1949). Wald proposed a unifying framework for much of the existing statistical theory, based on treating statistical inference as a special case of game theory, in which the decision maker faces nature in a zero-sum game. This leads to maximizing the smallest utility, rather than a subjective expectation of utility. The contrast of these two perspectives will continue through the next two chapters.

**Admissibility.** Admissibility is the most basic and influential rationality requirement of Wald's classical statistical decision theory. A nice surprise for Savage's fans is that maximizing expected utility is a safe way, and often, at least approximately, the only way, to build admissible statistical decision rules. Nice. In this chapter we also reinterpret one of the milestones of statistical theory, the Neyman–Pearson lemma (Neyman and Pearson 1933), in the light of the far-reaching theory this lemma sparked.

**Shrinkage.** The second major surprise from the study of admissibility is the fact that  $\bar{x}$ —the motherhood and apple pie of the statistical world—is inadmissible in estimating the mean of a multidimensional normal vector of observations. Stein (1955) was the first to realize this. We explore some of the important research directions

that stemmed from Stein's paper, including shrinkage estimation, empirical Bayes estimation, and hierarchical modeling.

**Scoring rules.** We change our focus to prediction and explore the implications of holding forecasters accountable for their predictions. We study the incentive systems that must be set in place for the forecasters to reveal their information/beliefs rather than using them to game the system. This leads to the study of proper scoring rules (Brier 1950). We also define and investigate calibration and refinement of forecasters.

**Choosing models** We try to understand whether statistical decision theory can be applied successfully to the much more elusive tasks of constructing and assessing statistical models. The jury is still out. On this puzzling note we close our tour of statistical decision theory and move to experimental design.

**Dynamic programming.** We describe a general approach for making decisions dynamically, so that we can both learn from accruing knowledge and plan ahead to account for how present decisions will affect future decisions and future knowledge. This approach, called dynamic programming, was developed by Bellman (1957). We will try to understand why the problem is so hard (the "curse of dimensionality").

**Changes in utility as information.** In decision theory, the value of the information carried by a data set depends on what we intend to do with the data once we have collected them. We use decision trees to quantify this value (DeGroot 1984). We also explore in more detail a specific way of measuring the information in a data set, which tries to capture "generic learning" rather than specific usefulness in a given problem (Lindley 1956).

**Sample size.** We finally come to terms with the single most common decision statisticians make in their daily activities: how big should a data set be? We try to understand how all the machinery we have been setting in place can help us and give some examples. Our discussion is based on the first complete formalization of Bayesian decision-theoretic approaches to sample size determination (Raiffa and Schlaifer 1961).

**Stopping.** Lastly, we apply dynamic programming to sequential data collection, where we have the option to stop an experiment after each observation. We discuss the stopping rule principle, which states that within the expected utility paradigm, the rule used to arrive at the decision to stop at a certain stage is not informative about parameters controlling the data-generating mechanism. We also study whether it is possible to design stopping rules that will stop experimentation only when one's favorite conclusion is reached.

A terrific preparatory reading for this book is Lindley (2000) who lays out the philosophy of Bayesian statistics in simple, concise, and compelling terms. As you progress through the book you will find, generally in each chapter's preamble, alternative texts that dwell on individual topics in greater depth than we do. Some are also listed next. A large number of textbooks overlap with ours and we make no attempt at being comprehensive. An early treatment of statistical decision theory is Raiffa and Schlaifer