

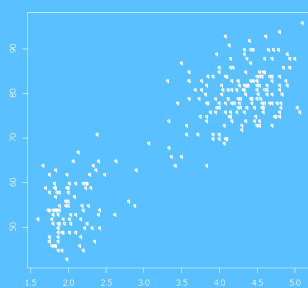
FOURTH EDITION

PETER M. LEE

BAYESIAN STATISTICS

An Introduction

 WILEY



Bayesian Statistics

Bayesian Statistics

An Introduction

Fourth Edition

Peter M. Lee

*Formerly Provost of Wentworth College,
University of York, UK*



A John Wiley & Sons, Ltd., Publication

This edition first published 2012
© 2012 John Wiley and 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.

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

Lee, Peter M.

Bayesian statistics : an introduction / Peter M. Lee. – 4th ed.

Includes bibliographical references and index.

ISBN 978-1-118-33257-3 (pbk.)

1. Bayesian statistical decision theory. I. Title.

QA279.5.L44 2012

519.5'42–dc23

2012007007

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

ISBN: 9781118332573

Typeset in 10/12pt Times by Aptara Inc., New Delhi, India

To The Memory of My Mother and of My Father

Contents

Preface	xix
Preface to the First Edition	xxi
1 Preliminaries	1
1.1 Probability and Bayes' Theorem	1
1.1.1 Notation	1
1.1.2 Axioms for probability	2
1.1.3 'Unconditional' probability	5
1.1.4 Odds	6
1.1.5 Independence	7
1.1.6 Some simple consequences of the axioms; Bayes' Theorem	7
1.2 Examples on Bayes' Theorem	9
1.2.1 The Biology of Twins	9
1.2.2 A political example	10
1.2.3 A warning	10
1.3 Random variables	12
1.3.1 Discrete random variables	12
1.3.2 The binomial distribution	13
1.3.3 Continuous random variables	14
1.3.4 The normal distribution	16
1.3.5 Mixed random variables	17
1.4 Several random variables	17
1.4.1 Two discrete random variables	17
1.4.2 Two continuous random variables	18
1.4.3 Bayes' Theorem for random variables	20
1.4.4 Example	21
1.4.5 One discrete variable and one continuous variable	21
1.4.6 Independent random variables	22
1.5 Means and variances	23
1.5.1 Expectations	23
1.5.2 The expectation of a sum and of a product	24
1.5.3 Variance, precision and standard deviation	25

1.5.4	Examples	25
1.5.5	Variance of a sum; covariance and correlation	27
1.5.6	Approximations to the mean and variance of a function of a random variable	28
1.5.7	Conditional expectations and variances	29
1.5.8	Medians and modes	31
1.6	Exercises on Chapter 1	31
2	Bayesian inference for the normal distribution	36
2.1	Nature of Bayesian inference	36
2.1.1	Preliminary remarks	36
2.1.2	Post is prior times likelihood	36
2.1.3	Likelihood can be multiplied by any constant	38
2.1.4	Sequential use of Bayes' Theorem	38
2.1.5	The predictive distribution	39
2.1.6	A warning	39
2.2	Normal prior and likelihood	40
2.2.1	Posterior from a normal prior and likelihood	40
2.2.2	Example	42
2.2.3	Predictive distribution	43
2.2.4	The nature of the assumptions made	44
2.3	Several normal observations with a normal prior	44
2.3.1	Posterior distribution	44
2.3.2	Example	46
2.3.3	Predictive distribution	47
2.3.4	Robustness	47
2.4	Dominant likelihoods	48
2.4.1	Improper priors	48
2.4.2	Approximation of proper priors by improper priors	49
2.5	Locally uniform priors	50
2.5.1	Bayes' postulate	50
2.5.2	Data translated likelihoods	52
2.5.3	Transformation of unknown parameters	52
2.6	Highest density regions	54
2.6.1	Need for summaries of posterior information	54
2.6.2	Relation to classical statistics	55
2.7	Normal variance	55
2.7.1	A suitable prior for the normal variance	55
2.7.2	Reference prior for the normal variance	58
2.8	HDRs for the normal variance	59
2.8.1	What distribution should we be considering?	59
2.8.2	Example	59
2.9	The role of sufficiency	60
2.9.1	Definition of sufficiency	60
2.9.2	Neyman's factorization theorem	61

2.9.3	Sufficiency principle	63
2.9.4	Examples	63
2.9.5	Order statistics and minimal sufficient statistics	65
2.9.6	Examples on minimal sufficiency	66
2.10	Conjugate prior distributions	67
2.10.1	Definition and difficulties	67
2.10.2	Examples	68
2.10.3	Mixtures of conjugate densities	69
2.10.4	Is your prior really conjugate?	71
2.11	The exponential family	71
2.11.1	Definition	71
2.11.2	Examples	72
2.11.3	Conjugate densities	72
2.11.4	Two-parameter exponential family	73
2.12	Normal mean and variance both unknown	73
2.12.1	Formulation of the problem	73
2.12.2	Marginal distribution of the mean	75
2.12.3	Example of the posterior density for the mean	76
2.12.4	Marginal distribution of the variance	77
2.12.5	Example of the posterior density of the variance	77
2.12.6	Conditional density of the mean for given variance	77
2.13	Conjugate joint prior for the normal distribution	78
2.13.1	The form of the conjugate prior	78
2.13.2	Derivation of the posterior	80
2.13.3	Example	81
2.13.4	Concluding remarks	82
2.14	Exercises on Chapter 2	82
3	Some other common distributions	85
3.1	The binomial distribution	85
3.1.1	Conjugate prior	85
3.1.2	Odds and log-odds	88
3.1.3	Highest density regions	90
3.1.4	Example	91
3.1.5	Predictive distribution	92
3.2	Reference prior for the binomial likelihood	92
3.2.1	Bayes' postulate	92
3.2.2	Haldane's prior	93
3.2.3	The arc-sine distribution	94
3.2.4	Conclusion	95
3.3	Jeffreys' rule	96
3.3.1	Fisher's information	96
3.3.2	The information from several observations	97
3.3.3	Jeffreys' prior	98

3.3.4	Examples	98
3.3.5	Warning	100
3.3.6	Several unknown parameters	100
3.3.7	Example	101
3.4	The Poisson distribution	102
3.4.1	Conjugate prior	102
3.4.2	Reference prior	103
3.4.3	Example	104
3.4.4	Predictive distribution	104
3.5	The uniform distribution	106
3.5.1	Preliminary definitions	106
3.5.2	Uniform distribution with a fixed lower endpoint	107
3.5.3	The general uniform distribution	108
3.5.4	Examples	110
3.6	Reference prior for the uniform distribution	110
3.6.1	Lower limit of the interval fixed	110
3.6.2	Example	111
3.6.3	Both limits unknown	111
3.7	The tramcar problem	113
3.7.1	The discrete uniform distribution	113
3.8	The first digit problem; invariant priors	114
3.8.1	A prior in search of an explanation	114
3.8.2	The problem	114
3.8.3	A solution	115
3.8.4	Haar priors	117
3.9	The circular normal distribution	117
3.9.1	Distributions on the circle	117
3.9.2	Example	119
3.9.3	Construction of an HDR by numerical integration	120
3.9.4	Remarks	122
3.10	Approximations based on the likelihood	122
3.10.1	Maximum likelihood	122
3.10.2	Iterative methods	123
3.10.3	Approximation to the posterior density	123
3.10.4	Examples	124
3.10.5	Extension to more than one parameter	126
3.10.6	Example	127
3.11	Reference posterior distributions	128
3.11.1	The information provided by an experiment	128
3.11.2	Reference priors under asymptotic normality	130
3.11.3	Uniform distribution of unit length	131
3.11.4	Normal mean and variance	132
3.11.5	Technical complications	134
3.12	Exercises on Chapter 3	134

4 Hypothesis testing	138
4.1 Hypothesis testing	138
4.1.1 Introduction	138
4.1.2 Classical hypothesis testing	138
4.1.3 Difficulties with the classical approach	139
4.1.4 The Bayesian approach	140
4.1.5 Example	142
4.1.6 Comment	143
4.2 One-sided hypothesis tests	143
4.2.1 Definition	143
4.2.2 P -values	144
4.3 Lindley's method	145
4.3.1 A compromise with classical statistics	145
4.3.2 Example	145
4.3.3 Discussion	146
4.4 Point (or sharp) null hypotheses with prior information	146
4.4.1 When are point null hypotheses reasonable?	146
4.4.2 A case of nearly constant likelihood	147
4.4.3 The Bayesian method for point null hypotheses	148
4.4.4 Sufficient statistics	149
4.5 Point null hypotheses for the normal distribution	150
4.5.1 Calculation of the Bayes' factor	150
4.5.2 Numerical examples	151
4.5.3 Lindley's paradox	152
4.5.4 A bound which does not depend on the prior distribution	154
4.5.5 The case of an unknown variance	155
4.6 The Doogian philosophy	157
4.6.1 Description of the method	157
4.6.2 Numerical example	157
4.7 Exercises on Chapter 4	158
5 Two-sample problems	162
5.1 Two-sample problems – both variances unknown	162
5.1.1 The problem of two normal samples	162
5.1.2 Paired comparisons	162
5.1.3 Example of a paired comparison problem	163
5.1.4 The case where both variances are known	163
5.1.5 Example	164
5.1.6 Non-trivial prior information	165
5.2 Variances unknown but equal	165
5.2.1 Solution using reference priors	165
5.2.2 Example	167
5.2.3 Non-trivial prior information	167

5.3	Variances unknown and unequal (Behrens–Fisher problem)	168
5.3.1	Formulation of the problem	168
5.3.2	Patil’s approximation	169
5.3.3	Example	170
5.3.4	Substantial prior information	170
5.4	The Behrens–Fisher controversy	171
5.4.1	The Behrens–Fisher problem from a classical standpoint	171
5.4.2	Example	172
5.4.3	The controversy	173
5.5	Inferences concerning a variance ratio	173
5.5.1	Statement of the problem	173
5.5.2	Derivation of the F distribution	174
5.5.3	Example	175
5.6	Comparison of two proportions; the 2×2 table	176
5.6.1	Methods based on the log-odds ratio	176
5.6.2	Example	177
5.6.3	The inverse root-sine transformation	178
5.6.4	Other methods	178
5.7	Exercises on Chapter 5	179
6	Correlation, regression and the analysis of variance	182
6.1	Theory of the correlation coefficient	182
6.1.1	Definitions	182
6.1.2	Approximate posterior distribution of the correlation coefficient	184
6.1.3	The hyperbolic tangent substitution	186
6.1.4	Reference prior	188
6.1.5	Incorporation of prior information	189
6.2	Examples on the use of the correlation coefficient	189
6.2.1	Use of the hyperbolic tangent transformation	189
6.2.2	Combination of several correlation coefficients	189
6.2.3	The squared correlation coefficient	190
6.3	Regression and the bivariate normal model	190
6.3.1	The model	190
6.3.2	Bivariate linear regression	191
6.3.3	Example	193
6.3.4	Case of known variance	194
6.3.5	The mean value at a given value of the explanatory variable	194
6.3.6	Prediction of observations at a given value of the explanatory variable	195
6.3.7	Continuation of the example	195
6.3.8	Multiple regression	196
6.3.9	Polynomial regression	196

6.4	Conjugate prior for the bivariate regression model	197
6.4.1	The problem of updating a regression line	197
6.4.2	Formulae for recursive construction of a regression line	197
6.4.3	Finding an appropriate prior	199
6.5	Comparison of several means – the one way model	200
6.5.1	Description of the one way layout	200
6.5.2	Integration over the nuisance parameters	201
6.5.3	Derivation of the F distribution	203
6.5.4	Relationship to the analysis of variance	203
6.5.5	Example	204
6.5.6	Relationship to a simple linear regression model	206
6.5.7	Investigation of contrasts	207
6.6	The two way layout	209
6.6.1	Notation	209
6.6.2	Marginal posterior distributions	210
6.6.3	Analysis of variance	212
6.7	The general linear model	212
6.7.1	Formulation of the general linear model	212
6.7.2	Derivation of the posterior	214
6.7.3	Inference for a subset of the parameters	215
6.7.4	Application to bivariate linear regression	216
6.8	Exercises on Chapter 6	217
7	Other topics	221
7.1	The likelihood principle	221
7.1.1	Introduction	221
7.1.2	The conditionality principle	222
7.1.3	The sufficiency principle	223
7.1.4	The likelihood principle	223
7.1.5	Discussion	225
7.2	The stopping rule principle	226
7.2.1	Definitions	226
7.2.2	Examples	226
7.2.3	The stopping rule principle	227
7.2.4	Discussion	228
7.3	Informative stopping rules	229
7.3.1	An example on capture and recapture of fish	229
7.3.2	Choice of prior and derivation of posterior	230
7.3.3	The maximum likelihood estimator	231
7.3.4	Numerical example	231
7.4	The likelihood principle and reference priors	232
7.4.1	The case of Bernoulli trials and its general implications	232
7.4.2	Conclusion	233

7.5	Bayesian decision theory	234
7.5.1	The elements of game theory	234
7.5.2	Point estimators resulting from quadratic loss	236
7.5.3	Particular cases of quadratic loss	237
7.5.4	Weighted quadratic loss	238
7.5.5	Absolute error loss	238
7.5.6	Zero-one loss	239
7.5.7	General discussion of point estimation	240
7.6	Bayes linear methods	240
7.6.1	Methodology	240
7.6.2	Some simple examples	241
7.6.3	Extensions	243
7.7	Decision theory and hypothesis testing	243
7.7.1	Relationship between decision theory and classical hypothesis testing	243
7.7.2	Composite hypotheses	245
7.8	Empirical Bayes methods	245
7.8.1	Von Mises' example	245
7.8.2	The Poisson case	246
7.9	Exercises on Chapter 7	247
8	Hierarchical models	253
8.1	The idea of a hierarchical model	253
8.1.1	Definition	253
8.1.2	Examples	254
8.1.3	Objectives of a hierarchical analysis	257
8.1.4	More on empirical Bayes methods	257
8.2	The hierarchical normal model	258
8.2.1	The model	258
8.2.2	The Bayesian analysis for known overall mean	259
8.2.3	The empirical Bayes approach	261
8.3	The baseball example	262
8.4	The Stein estimator	264
8.4.1	Evaluation of the risk of the James–Stein estimator	267
8.5	Bayesian analysis for an unknown overall mean	268
8.5.1	Derivation of the posterior	270
8.6	The general linear model revisited	272
8.6.1	An informative prior for the general linear model	272
8.6.2	Ridge regression	274
8.6.3	A further stage to the general linear model	275
8.6.4	The one way model	276
8.6.5	Posterior variances of the estimators	277
8.7	Exercises on Chapter 8	277

9	The Gibbs sampler and other numerical methods	281
9.1	Introduction to numerical methods	281
9.1.1	Monte Carlo methods	281
9.1.2	Markov chains	282
9.2	The <i>EM</i> algorithm	283
9.2.1	The idea of the <i>EM</i> algorithm	283
9.2.2	Why the <i>EM</i> algorithm works	285
9.2.3	Semi-conjugate prior with a normal likelihood	287
9.2.4	The <i>EM</i> algorithm for the hierarchical normal model	288
9.2.5	A particular case of the hierarchical normal model	290
9.3	Data augmentation by Monte Carlo	291
9.3.1	The genetic linkage example revisited	291
9.3.2	Use of R	291
9.3.3	The genetic linkage example in R	292
9.3.4	Other possible uses for data augmentation	293
9.4	The Gibbs sampler	294
9.4.1	Chained data augmentation	294
9.4.2	An example with observed data	296
9.4.3	More on the semi-conjugate prior with a normal likelihood	299
9.4.4	The Gibbs sampler as an extension of chained data augmentation	301
9.4.5	An application to change-point analysis	302
9.4.6	Other uses of the Gibbs sampler	306
9.4.7	More about convergence	309
9.5	Rejection sampling	311
9.5.1	Description	311
9.5.2	Example	311
9.5.3	Rejection sampling for log-concave distributions	311
9.5.4	A practical example	313
9.6	The Metropolis–Hastings algorithm	317
9.6.1	Finding an invariant distribution	317
9.6.2	The Metropolis–Hastings algorithm	318
9.6.3	Choice of a candidate density	320
9.6.4	Example	321
9.6.5	More realistic examples	322
9.6.6	Gibbs as a special case of Metropolis–Hastings	322
9.6.7	Metropolis within Gibbs	323
9.7	Introduction to WinBUGS and OpenBUGS	323
9.7.1	Information about WinBUGS and OpenBUGS	323
9.7.2	Distributions in WinBUGS and OpenBUGS	324
9.7.3	A simple example using WinBUGS	324
9.7.4	The pump failure example revisited	327
9.7.5	DoodleBUGS	327

9.7.6	coda	329
9.7.7	R2WinBUGS and R2OpenBUGS	329
9.8	Generalized linear models	332
9.8.1	Logistic regression	332
9.8.2	A general framework	334
9.9	Exercises on Chapter 9	335
10	Some approximate methods	340
10.1	Bayesian importance sampling	340
10.1.1	Importance sampling to find HDRs	343
10.1.2	Sampling importance re-sampling	344
10.1.3	Multidimensional applications	344
10.2	Variational Bayesian methods: simple case	345
10.2.1	Independent parameters	347
10.2.2	Application to the normal distribution	349
10.2.3	Updating the mean	350
10.2.4	Updating the variance	351
10.2.5	Iteration	352
10.2.6	Numerical example	352
10.3	Variational Bayesian methods: general case	353
10.3.1	A mixture of multivariate normals	353
10.4	<i>ABC</i> : Approximate Bayesian Computation	356
10.4.1	The <i>ABC</i> rejection algorithm	356
10.4.2	The genetic linkage example	358
10.4.3	The <i>ABC</i> Markov Chain Monte Carlo algorithm	360
10.4.4	The <i>ABC</i> Sequential Monte Carlo algorithm	362
10.4.5	The <i>ABC</i> local linear regression algorithm	365
10.4.6	Other variants of <i>ABC</i>	366
10.5	Reversible jump Markov chain Monte Carlo	367
10.5.1	<i>RJMCMC</i> algorithm	367
10.6	Exercises on Chapter 10	369
Appendix A	Common statistical distributions	373
A.1	Normal distribution	374
A.2	Chi-squared distribution	375
A.3	Normal approximation to chi-squared	376
A.4	Gamma distribution	376
A.5	Inverse chi-squared distribution	377
A.6	Inverse chi distribution	378
A.7	Log chi-squared distribution	379
A.8	Student's <i>t</i> distribution	380
A.9	Normal/chi-squared distribution	381
A.10	Beta distribution	382
A.11	Binomial distribution	383
A.12	Poisson distribution	384

A.13	Negative binomial distribution	385
A.14	Hypergeometric distribution	386
A.15	Uniform distribution	387
A.16	Pareto distribution	388
A.17	Circular normal distribution	389
A.18	Behrens' distribution	391
A.19	Snedecor's F distribution	393
A.20	Fisher's z distribution	393
A.21	Cauchy distribution	394
A.22	The probability that one beta variable is greater than another	395
A.23	Bivariate normal distribution	395
A.24	Multivariate normal distribution	396
A.25	Distribution of the correlation coefficient	397
Appendix B	Tables	399
B.1	Percentage points of the Behrens–Fisher distribution	399
B.2	Highest density regions for the chi-squared distribution	402
B.3	HDRs for the inverse chi-squared distribution	404
B.4	Chi-squared corresponding to HDRs for log chi-squared	406
B.5	Values of F corresponding to HDRs for log F	408
Appendix C	R programs	430
Appendix D	Further reading	436
D.1	Robustness	436
D.2	Nonparametric methods	436
D.3	Multivariate estimation	436
D.4	Time series and forecasting	437
D.5	Sequential methods	437
D.6	Numerical methods	437
D.7	Bayesian networks	437
D.8	General reading	438
References		439
Index		455

Note: The tables in the Appendix are intended for use in conjunction with a standard set of statistical tables, for example, Lindley and Scott (1995) or Neave (1978). They extend the coverage of these tables so that they are roughly comparable with those of Isaacs *et al.* (1974) or with the tables in Appendix A of Novick and Jackson (1974). However, tables of values easily computed with a pocket calculator have been omitted. The tables have been computed using NAG routines and algorithms described in Patil (1965) and Jackson (1974).

Preface

When I started writing this book in 1987 it never occurred to me that it would still be of interest nearly a quarter of a century later, but it appears that it is, and I am delighted to introduce a fourth edition. The subject moves ever onwards, with increasing emphasis on Monte-Carlo based techniques. With this in mind, Chapter 9 entitled ‘The Gibbs sampler’ has been considerably extended (including more numerical examples and treatments of OpenBUGS, R2WinBUGS and R2OpenBUGS) and a new Chapter 10 covering Bayesian importance sampling, variational Bayes, ABC (Approximate Bayesian Computation) and RJMCMC (Reversible Jump Markov Chain Monte Carlo) has been added. Mistakes and misprints in the third edition have been corrected and minor alterations made throughout.

The basic idea of using Bayesian methods has become more and more popular, and a useful accessible account for the layman has been written by McGrayne (2011). There is every reason to believe that an approach to statistics which I began teaching in 1985 with some misgivings because of its unfashionability will continue to gain adherents. The fact is that the Bayesian approach produces results in a comprehensible form and with modern computational methods produces them quickly and easily.

Useful comments for which I am grateful were received from John Burkett, Stephen Connor, Jacco Thijssen, Bo Wang and others; they, of course, have no responsibility for any deficiencies in the end result.

The website associated with the book

<http://www-users.york.ac.uk/~pml1/bayes/book.htm>

(note that in the above pml are letters followed by the digit 1) works through all the numerical examples in R as well as giving solutions to all the exercises in the book (and some further exercises to which the solutions are not given).

Peter M. Lee
19 December 2011

Preface to the First Edition

When I first learned a little statistics, I felt confused, and others I spoke to confessed that they had similar feelings. Not because the mathematics was difficult – most of that was a lot easier than pure mathematics – but because I found it difficult to follow the logic by which inferences were arrived at from data. It sounded as if the statement that a null hypothesis was rejected at the 5% level meant that there was only a 5% chance of that hypothesis was true, and yet the books warned me that this was not a permissible interpretation. Similarly, the statement that a 95% confidence interval for an unknown parameter ran from -2 to $+2$ sounded as if the parameter lay in that interval with 95% probability and yet I was warned that all I could say was that if I carried out similar procedures time after time then the unknown parameters would lie in the confidence intervals I constructed 95% of the time. It appeared that the books I looked at were not answering the questions that would naturally occur to a beginner, and that instead they answered some rather recondite questions which no one was likely to want to ask.

Subsequently, I discovered that the whole theory had been worked out in very considerable detail in such books as Lehmann (1986). But attempts such as those that Lehmann describes to put everything on a firm foundation raised even more questions. I gathered that the usual t test could be justified as a procedure that was ‘uniformly most powerful unbiased’, but I could only marvel at the ingenuity that led to the invention of such criteria for the justification of the procedure, while remaining unconvinced that they had anything sensible to say about a general theory of statistical inference. Of course Lehmann and others with an equal degree of common sense were capable of developing more and more complicated constructions and exceptions so as to build up a theory that appeared to cover most problems without doing anything obviously silly, and yet the whole enterprise seemed reminiscent of the construction of epicycle upon epicycle in order to preserve a theory of planetary motion based on circular motion; there seemed to be an awful lot of ‘ad hocery’.

I was told that there was another theory of statistical inference, based ultimately on the work of the Rev. Thomas Bayes, a Presbyterian minister, who lived from 1702 to 1761 whose key paper was published posthumously by his friend Richard Price as Bayes (1763) [more information about Bayes himself and his work can be

found in Holland (1962), Todhunter (1865, 1949) and Stigler (1986a)].¹ However, I was warned that there was something not quite proper about this theory, because it depended on your personal beliefs and so was not objective. More precisely, it depended on taking some expression of your beliefs about an unknown quantity before the data was available (your ‘prior probabilities’) and modifying them in the light of the data (via the so-called ‘likelihood function’) to arrive at your ‘posterior probabilities’ using the formulation that ‘posterior is proportional to prior times likelihood’. The standard, or ‘classical’, theory of statistical inference, on the other hand, was said to be objective, because it does not refer to anything corresponding to the Bayesian notion of ‘prior beliefs’. Of course, the fact that in this theory, you sometimes looked for a 5% significance test and sometimes for a 0.1% significance test, depending on what you thought about the different situations involved, was said to be quite a different matter.

I went on to discover that this theory could lead to the sorts of conclusions that I had naïvely expected to get from statistics when I first learned about it. Indeed, some lecture notes of Lindley’s [and subsequently his book, Lindley (1965)] and the pioneering book by Jeffreys (1961) showed that if the statistician had ‘personal probabilities’ that were of a certain conventional type then conclusions very like those in the elementary books I had first looked at could be arrived at, with the difference that a 95% confidence interval really did mean an interval in which the statistician was justified in thinking that there was a 95% probability of finding the unknown parameter. On the other hand, there was the further freedom to adopt other initial choices of personal beliefs and thus to arrive at different conclusions.

Over a number of years I taught the standard, classical, theory of statistics to a large number of students, most of whom appeared to have similar difficulties to those I had myself encountered in understanding the nature of the conclusions that this theory comes to. However, the mere fact that students have difficulty with a theory does not prove it wrong. More importantly, I found the theory did not improve with better acquaintance, and I went on studying Bayesian theory. It turned out that there were real differences in the conclusions arrived at by classical and Bayesian statisticians, and so the former was not just a special case of the latter corresponding to a conventional choice of prior beliefs. On the contrary, there was a strong disagreement between statisticians as to the conclusions to be arrived at in certain standard situations, of which I will cite three examples for now. One concerns a test of a sharp null hypothesis (e.g. a test that the mean of a distribution is exactly equal to zero), especially when the sample size was large. A second concerns the Behrens–Fisher problem, that is, the inferences that can be made about the difference between the means of two populations when no assumption is made about their variances. Another is the likelihood principle, which asserts that you can only take account of the probability of events that have actually occurred under various hypotheses, and not of events that might have happened but did not; this

¹ Further information is now available in Bellhouse (2003) and Dale (2003). Useful information can also be found in Bellhouse *et al.* (1988–1992), Dale (1999), Edwards (1993, 2004) and Hald (1986, 1998, 2007).

principle follows from Bayesian statistics and is contradicted by the classical theory. A particular case concerns the relevance of stopping rules, that is to say whether or not you are entitled to take into account the fact that the experimenter decided when to stop experimenting depending on the results so far available rather than having decided to use a fixed sample size all along. The more I thought about all these controversies, the more I was convinced that the Bayesians were right on these disputed issues.

At long last, I decided to teach a third-year course on Bayesian statistics in the University of York, which I have now done for a few years. Most of the students who took the course did find the theory more coherent than the classical theory they had learned in the first course on mathematical statistics they had taken in their second year, and I became yet more clear in my own mind that this was the right way to view statistics. I do, however, admit that there are topics (such as non-parametric statistics) which are difficult to fit into a Bayesian framework.

A particular difficulty in teaching this course was the absence of a suitable book for students who were reasonably well prepared mathematically and already knew some statistics, even if they knew nothing of Bayes apart from Bayes' theorem. I wanted to teach them more, and to give more information about the incorporation of real as opposed to conventional prior information, than they could get from Lindley (1965), but I did not think they were well enough prepared to face books like Box and Tiao (1973) or Berger (1985), and so I found that in teaching the course I had to get together material from a large number of sources, and in the end found myself writing this book. It seems less and less likely that students in mathematics departments will be completely unfamiliar with the ideas of statistics, and yet they are not (so far) likely to have encountered Bayesian methods in their first course on statistics, and this book is designed with these facts in mind. It is assumed that the reader has a knowledge of calculus of one and two variables and a fair degree of mathematical maturity, but most of the book does not assume a knowledge of linear algebra. The development of the text is self-contained, but from time to time the contrast between Bayesian and classical conclusions is pointed out, and it is supposed that in most cases the reader will have some idea as to the conclusion that a classical statistician would come to, although no very detailed knowledge of classical statistics is expected. It should be possible to use the book as a course text for final year undergraduate or beginning graduate students or for self-study for those who want a concise account of the way in which the Bayesian approach to statistics develops and the contrast between it and the conventional approach. The theory is built up step by step, rather than doing everything in the greatest generality to start with, and important notions such as sufficiency are brought out of a discussion of the salient features of specific examples.

I am indebted to Professor RA Cooper for helpful comments on an earlier draft of this book, although of course he cannot be held responsible for any errors in the final version.

Peter M. Lee
30 March 1988

1

Preliminaries

1.1 Probability and Bayes' Theorem

1.1.1 Notation

The notation will be kept simple as possible, but it is useful to express statements about probability in the language of set theory. You probably know most of the symbols undermentioned, but if you do not you will find it easy enough to get the hang of this useful shorthand. We consider sets $A, B, C, \dots$ of elements $x, y, z, \dots$ and we use the word 'iff' to mean 'if and only if'. Then we write

$x \in A$ iff x is a member of A ;

$x \notin A$ iff x is *not* a member of A ;

$A = \{x, y, z\}$ iff A is the set whose only members are x, y and z (and similarly for larger or smaller sets);

$A = \{x; S(x)\}$ iff A is the set of elements for which the statement $S(x)$ is true;

$\emptyset = \{x; x \neq x\}$ for the null set, that is the set with no elements;

$x \notin \emptyset$ for all x ;

$A \subset B$ (i.e. A is a subset of B) iff $x \in A$ implies $x \in B$;

$A \supset B$ (i.e. A is a superset of B) iff $x \in A$ is implied by $x \in B$;

$\emptyset \subset A, A \subset A$ and $A \supset A$ for all A ;

$A \cup B = \{x; x \in A \text{ or } x \in B\}$ (where ' P or Q ' means ' P or Q or both') (referred to as the union of A and B or as A union B);

$AB = A \cap B = \{x; x \in A \text{ and } x \in B\}$ (referred to as the intersection of A and B or as A intersect B);

A and B are disjoint iff $AB = \emptyset$;

$A \setminus B = \{x; x \in A, \text{ but } x \notin B\}$ (referred to as the difference set A less B).

Let (A_n) be a sequence $A_1, A_2, A_3, \dots$ of sets. Then

$$\bigcup_{n=1}^{\infty} A_n = \{x; x \in A_n \text{ for one or more } n\};$$

$$\bigcap_{n=1}^{\infty} A_n = \{x; x \in A_n \text{ for all } n\};$$

$$(A_n) \text{ exhausts } B \text{ if } \bigcup_{i=1}^{\infty} A_n \supset B;$$

$$(A_n) \text{ consists of exclusive sets if } A_m A_n = \emptyset \text{ for } m \neq n;$$

$$(A_n) \text{ consists of exclusive sets given } B \text{ if } A_m A_n B = \emptyset \text{ for } m \neq n;$$

$$(A_n) \text{ is non-decreasing if } A_1 \subset A_2 \subset \dots, \text{ that is } A_n \subset A_{n+1} \text{ for all } n;$$

$$(A_n) \text{ is non-increasing if } A_1 \supset A_2 \supset \dots, \text{ that is } A_n \supset A_{n+1} \text{ for all } n.$$

We sometimes need a notation for intervals on the real line, namely

$$[a, b] = \{x; a \leq x \leq b\};$$

$$(a, b) = \{x; a < x < b\};$$

$$[a, b) = \{x; a \leq x < b\};$$

$$(a, b] = \{x; a < x \leq b\}$$

where a and b are real numbers or $+\infty$ or $-\infty$.

1.1.2 Axioms for probability

In the study of probability and statistics, we refer to as complete a description of the situation as we need in a particular context as an *elementary event*.

Thus, if we are concerned with the tossing of a red die and a blue die, then a typical elementary event is ‘red three, blue five’, or if we are concerned with the numbers of Labour and Conservative MPs in the next parliament, a typical elementary event is ‘Labour 350, Conservative 250’. Often, however, we want to talk about one aspect of the situation. Thus, in the case of the first example, we might be interested in whether or not we get a red three, which possibility includes ‘red three, blue one’, ‘red three, blue two’, etc. Similarly, in the other example, we could be interested in whether there is a Labour majority of at least 100, which can also be analyzed into elementary events. With this in mind, an *event* is defined as a set of elementary events (this has the slightly curious consequence that, if you are *very* pedantic, an elementary event is not an event since it is an element rather than a set). We find it useful to say that one event E implies another event F if E is contained in F . Sometimes it is useful to generalize this by saying that, given H , E implies F if EH is contained in F . For example, given a red three has been thrown, throwing a blue three implies throwing an even total.

Note that the definition of an elementary event depends on the context. If we were never going to consider the blue die, then we could perfectly well treat events such as ‘red three’ as elementary events. In a particular context, the elementary events in terms of which it is sensible to work are usually clear enough.

Events are referred to above as possible future occurrences, but they can also describe present circumstances, known or unknown. Indeed, the relationship which probability attempts to describe is one between what you currently know and

something else about which you are uncertain, both of them being referred to as events. In other words, for at least some pairs of events E and H there is a number $P(E|H)$ defined which is called the probability of the event E given the hypothesis H . I might, for example, talk of the probability of the event E that I throw a red three given the hypothesis H that I have rolled two fair dice once, or the probability of the event E of a Labour majority of at least 100 given the hypothesis H which consists of my knowledge of the political situation to date. Note that in this context, the term ‘hypothesis’ can be applied to a large class of events, although later on we will find that in statistical arguments, we are usually concerned with hypotheses which are more like the hypotheses in the ordinary meaning of the word.

Various attempts have been made to define the notion of probability. Many early writers claimed that $P(E|H)$ was m/n where there were n symmetrical and so equally likely possibilities given H of which m resulted in the occurrence of E . Others have argued that $P(E|H)$ should be taken as the long run frequency with which E happens when H holds. These notions can help your intuition in some cases, but I think they are impossible to turn into precise, rigorous definitions. The difficulty with the first lies in finding genuinely ‘symmetrical’ possibilities – for example, real dice are only approximately symmetrical. In any case, there is a danger of circularity in the definitions of symmetry and probability. The difficulty with the second is that we never know how long we have to go on trying before we are within, say, 1% of the true value of the probability. Of course, we may be able to give a value for the number of trials we need to be within 1% of the true value with, say, probability 0.99, but this is leading to another vicious circle of definitions. Another difficulty is that sometimes we talk of the probability of events (e.g. nuclear war in the next 5 years) about which it is hard to believe in a large numbers of trials, some resulting in ‘success’ and some in ‘failure’. A good, brief discussion is to be found in Nagel (1939) and a fuller, more up-to-date one in Chatterjee (2003).

It seems to me, and to an increasing number of statisticians, that the only satisfactory way of thinking of $P(E|H)$ is as a measure of my degree of belief in E given that I know that H is true. It seems reasonable that this measure should abide by the following axioms:

- P1 $P(E|H) \geq 0$ for all E, H .
- P2 $P(H|H) = 1$ for all H .
- P3 $P(E \cup F|H) = P(E|H) + P(F|H)$ when $EFH = \emptyset$.
- P4 $P(E|FH)P(F|H) = P(EFH|H)$.

By taking $F = H \setminus E$ in P3 and using P1 and P2, it easily follows that

$$P(E|H) \leq 1 \text{ for all } E, H,$$

so that $P(E|H)$ is always between 0 and 1. Also by taking $F = \emptyset$ in P3 it follows that

$$P(\emptyset|H) = 0.$$

Now intuitive notions about probability always seem to agree that it should be a quantity between 0 and 1 which falls to 0 when we talk of the probability of something we are certain will not happen and rises to 1 when we are certain it will happen (and we are certain that H is true given H is true). Further, the additive property in P3 seems highly reasonable – we would, for example, expect the probability that the red die lands three or four should be the sum of the probability that it lands three and the probability that it lands four.

Axiom P4 may seem less familiar. It is sometimes written as

$$P(E|FH) = \frac{P(EF|H)}{P(F|H)}$$

although, of course, this form cannot be used if the denominator (and hence the numerator) on the right-hand side vanishes. To see that it is a reasonable thing to assume, consider the following data on criminality among the twin brothers or sisters of criminals [quoted in his famous book by Fisher (1925b)]. The twins were classified according as they had a criminal conviction (C) or not (N) and according as they were monozygotic (M) (which is more or less the same as identical – we will return to this in Section 1.2) or dizygotic (D), resulting in the following table:

	C	N	Total
M	10	3	13
D	2	15	17
Total	12	18	30

If we denote by H the knowledge that an individual has been picked at random from this population, then it seems reasonable to say that

$$\begin{aligned} P(C|H) &= 12/30, \\ P(MC|H) &= 10/30. \end{aligned}$$

If on the other hand, we consider an individual picked at random from among the twins with a criminal conviction in the population, we see that

$$P(M|CH) = 10/12$$

and hence

$$P(M|CH)P(C|H) = P(MC|H),$$

so that P4 holds in this case. It is easy to see that this relationship does not depend on the particular numbers that happen to appear in the data.

In many ways, the argument in the preceding paragraph is related to derivations of probabilities from symmetry considerations, so perhaps it should be stressed that