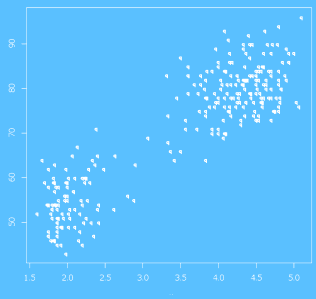
FOURTH EDITION

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

WILEY
A John Wiley & Sons, Ltd., Publication

This edition first published 2012

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

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

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 'adhockery'.

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 $A, B, C, \ldots$ of elements $x, y, z, \ldots$ 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 $AB = \emptyset$;

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

Bayesian Statistics: An Introduction, Fourth Edition. Peter M. Lee.

Let (A_n) be a sequence $A_1, A_2, A_3, \ldots$ 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) exhausts B if $\bigcup_{i=1}^{\infty} A_n \supset B$;
(A_n) consists of exclusive sets if $A_m A_n = \emptyset$ for $m \neq n$;
(A_n) consists of exclusive sets given B if $A_m A_n B = \emptyset$ for $m \neq n$;
(A_n) is non-decreasing if $A_1 \subset A_2 \subset \ldots$, that is $A_n \subset A_{n+1}$ for all n;
(A_n) is non-increasing if $A_1 \supset A_2 \supset \ldots$, that is $A_n \supset A_{n+1}$ for all n.

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

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

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 $\mathsf{P}(E|H)$ 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 $\mathsf{P}(E|H)$ 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 $\mathsf{P}(E|H)$ 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, rigourous 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 $\mathsf{P}(E|H)$ 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	$\mathsf{P}(E\|H)$	$\geqslant$	0	for all E, H.
P2	$\mathsf{P}(H\|H)$	$=$	1	for all H.
P3	$\mathsf{P}(E \cup F\|H)$	$=$	$\mathsf{P}(E\|H) + \mathsf{P}(F\|H)$	when $EFH = \emptyset$.
P4	$\mathsf{P}(E\|FH)\,\mathsf{P}(F\|H)$	$=$	$\mathsf{P}(EF\|H)$.	

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

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

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

$$\mathsf{P}(\emptyset|H) = 0.$$

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

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

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

$$\mathsf{P}(C|H) = 12/30,$$
$$\mathsf{P}(MC|H) = 10/30.$$

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

$$\mathsf{P}(M|CH) = 10/12$$

and hence

$$\mathsf{P}(M|CH)\mathsf{P}(C|H) = \mathsf{P}(MC|H),$$

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